TAMPERE UNIVERSITY OF TECHNOLOGY

JONI VUORIO

Binding of Hyaluronic Acid to Its CD44 Receptor

Master of Science Thesis

Examiners: Ilpo Vattulainen & Minna Kellomäki
Examiners and topic approved in
Science and Environmental Engineering
Faculty Council Meeting on 8.11.2013

TIIVISTELMÄ

TAMPEREEN TEKNILLINEN YLIOPISTO
Biotekniikan koulutusohjelma
JONI VUORIO: Hyaluronihapon sitoutuminen CD44-reseptoriinsa
Diplomityö, 87 sivua
Marraskuu 2013
Pääaine: Kudosteknologia
Tarkastajat: Prof. Ilpo Vattulainen & Prof. Minna Kellomäki
Avainsanat: Hyaluronihappoa sitova domeeni, molekyylidynamiikka, vapaaenergia, umbrella sampling -menetelmä, glykosylaatio

CD44 on solukalvon reseptoriproteiini, joka sitoutuu reversiibelisti hiilihydraattiligandiinsa, hyaluronihappoon. Tämä proteiini-hiilihydraattivuorovaikutus mahdollistaa sekä normaalin solumigraation, kuten valkosolujen pyörimisliikkeen, että pahanlaatuisten syöpäsolujen tunkeutumisen terveisiin kudoksiin muodostaessaan etäpesäkkeitä verenkierron välityksellä. Normaali solu tarvitsee näin ollen tehokkaita mekanismeja CD44:n ja hyaluronihapon vuorovaikutusaffiniteetin säätelyyn. Näiden säätelymekanismien tunnistaminen on osoittautunut kuitenkin haastavaksi. Perinteisiä laboratoriokokeita on vaikeuttanut käytettävien menetelmien heikko mittatarkkuus nanoskaalan ilmiöiden kuvaamisessa. Tietokonesimulaatioiden luotettavuutta ovat puolestaan rajoittaneet sekä simulaatiomallien epätarkkuudet että laskentaresurssien puute.

Tässä tutkimuksessa käytämme atomaarisen tarkkuuden omaavia molekyylidynamiikkasimulaatioita selvittämään, kuinka hyaluronihappo sitoutuu ihmisen villityypin CD44-proteiiniin. Tutkimus keskittyy erityisesti kolmeen potentiaaliseen säätelymekanismiin: hyaluronihapon pituuteen, proteiinin glykosylaatioihin ja proteiinin rakennemuutoksiin. Löydöksemme ovat varsin merkittäviä. Hyaluronihappo-oktameerin adsorptiolle lasketut vapaaenergiaprofiilit muun muassa osoittavat, että yksittäisen CD44-hyaluronihappovuorovaikutuksen voimakkuus on peräti 25 $\mathrm{kJ\,mol^{-1}}$. Lisäksi glykosyloimalla proteiinin aminohappotähteet Asn25, Asn100 ja Asn110 näytämme, että Asn25:een liitetty varausneutraali, viisi sokeriyksikköä sisältävä hiilihydraatti estää normaalit CD44-hyaluronihappovuorovaikutukset laskien sitoutumisen voimakkuutta 40 %. Kaiken tämän ohessa simulaatiodatamme osoittaa aiemmin vahvaan sitoutumiseen yhdistetyn konformaatiomuutoksen olevan itse asiassa molekylaarinen mekanismi, joka irrottaa proteiinin pintaan sitoutuneen hyaluronihappojuosteen.

Tutkimus paljastaa miten solu säätelee CD44:n ja hyaluronihapon sitoutumista toisiinsa. Uusi tieto voi edistää uudenlaisten lääkkeiden ja hoitomuotojen kehitystä ja kohdentamista esimerkiksi syövän hoitoon. Lisäksi tämän tutkimuksen tulokset hyödyttävät muiden vastaavien kalvoreseptoreiden proteiini-hiilihydraattivuorovaikutusten selvitystyötä.

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Programme in Biotechnology
JONI VUORIO : Binding of Hyaluronic Acid to its CD44 Receptor
Master of Science Thesis, 87 pages
November 2013
Major: Tissue Engineering
Examiners: Prof. Ilpo Vattulainen & Prof. Minna Kellomäki
Keywords: Hyaluronan binding domain, molecular dynamics, free energy, umbrella sampling, glycosylation

CD44 is a transmembrane receptor protein binding its carbohydrate ligand, hyaluronic acid (HA), in a reversible fashion. In addition to enabling normal cell migration, such as the rolling of white blood cells, these carbohydrate-protein interactions are exploited by malignant cancer cells metastasizing through the blood stream. A normal cell therefore requires effective regulatory mechanisms for controlling the binding affinity between CD44 and HA. Earlier studies addressing this topic have, however, been unable to identify these regulatory mechanisms. More precisely, traditional wet-lab experiments, limited by their spatial and temporal resolution, have been inadequate in describing transient nanoscale phenomena, such as the ligand binding. Previous computer simulations have, on the other hand, been limited by both veracity of simulation models and availability of computational resources.

In this Thesis, we use all-atom explicit-solvent molecular dynamics (MD) simulations to study the adsorption of HA oligomer to a human wild-type CD44 HA binding domain (HABD). In practice, we explore the role of three potential regulation mechanisms: size of the ligand, glycosylation of the protein, and conformation of the protein. First, free energy profiles for the adsorption of HA octamer, revealing the strength of an individual CD44-HA interaction to be over 25 $\mathrm{kJ\,mol^{-1}}$, suggest ligand binding to be irreversible. Second, by glycosylating the HABD at residues Asn25, Asn100, and Asn110, we show the first of these residues to block most of the native binding interactions. As a result, the strength of the adsorption is, at least when using charge-neutral core pentasaccharides as the attached carbohydrates, decreased by 40 %. More surprisingly, the simulation data reveal a conformation transition previously correlated with high-affinity binding to, in fact, act as a molecular mechanism repelling the bound HA oligomer, and thereby dynamically regulating the biological activity of the CD44.

The findings of this study uncover how the binding of HA to its CD44 receptor is regulated. This information may facilitate the design and targeting of novel drugs and therapies against conditions, such as cancer. Lastly, the insight from this study is of potential value when considering the carbohydrate-protein interactions of other cell surface receptors.

PREFACE

This Master of Science Thesis was written while working in the Biological Physics and Soft Matter (BIO) research group of the Department of Physics of Tampere University of Technology between June 2013 and November 2013. The molecular dynamics simulations were conducted using the computing services of the Finnish IT Centre for Scientific Computing (CSC) and the Tampere Center for Scientific Computing (TCSC).

First, I would like to express my gratitude to Prof. Ilpo Vattulainen for examining this thesis work, and especially for giving me the chance to work with such an excellent group of scientists. I also wish to thank my other examiner Prof. Minna Kellomäki for her collaboration. This project and this field in general, have managed to combine all my scientific interests, and thereby have led me to pursue a research career in biophysics as a post-graduate student.

I am also very thankful to the other members of this project. I especially want to thank my supervisor Ph.D. Hector Martinez-Seara Monné for the long and fruitful discussions we had through the project. A great thank also goes to Reinis Danne for his work with the topologies and to Kalle Koivuniemi for sharing some of the huge simulation load with me.

Still, I would like to thank the other members of our group, particularly all my past and present roommates in the original Batcave (SG315/317). It has been an inspiring and relaxed atmosphere to work in. I would also like to thank Matti Javanainen for his solid technical and scientific support. Additionally, special thanks go to Sanja Pöyry and Sami Rissanen for their helpful tips and corrections.

Lastly, I wish to thank my family and friends for their company and assistance. I am especially thankful to my mother who always so unselfishly offers me her greatly appreciated help and support.

CONTENTS

SYMBOLS & ABBREVIATIONS

A	Helmholz free energy
AMBER	Assisted Model Building with Energy Refinement, a force field
$\mathbf{b}$	Matrix for box vectors in Parrinello–Rahman pressure coupling
CS	Chondroitin sulphate
ECM	Extracellular matrix
ER	Endoplasmic reticulum
erfc	Complementary error function in Ewald summation
E_{kin}	Kinetic energy
$E_{\mathrm{kin},0}$	Target kinetic energy
$\mathbf{F}_f$	Free energy constant
$\mathbf{F}_i$	Force acting on particle i
f_{ij}	Scaling factor for non-bonded interactions
G	Gibbs free energy
GAG	Glycosaminoglycan
GlcNAc	*N*-acetyl-D-glucosamine
GalNAc	*N*-acetyl-D-galactosamine
GlcUA	D-glucuronic acid
GROMACS	GROningen MAchine for Chemical Simulations
$\mathcal{H}_{\mathrm{bonded}}$	Potential function for bonded interactions
$\mathcal{H}_{\mathrm{Ewald,Coulomb}}$	Coulombic potential given by Ewald sum
$\mathcal{H}_{\mathrm{Ewald,real}}$	Real space term of Ewald summation
$\mathcal{H}_{\mathrm{Ewald,recip}}$	Reciprocal space term of Ewald summation
$\mathcal{H}_{\mathrm{Ewald,corr}}$	Correction term of Ewald summation
$\mathcal{H}_{\mathrm{non-bonded}}$	Potential function for non-bonded interactions
$\mathcal{H}_{\mathrm{total}}$	Total potential function
HA	Hyaluronic acid or hyaluronan
HA_1	Hyaluronan fragment of one disaccharide unit
HA_2	Hyaluronan fragment of two disaccharide units
HA_8	Hyaluronan fragment of eight disaccharide units
HAS	Hyaluronan synthase enzyme
HS	Heparan sulphate
k_{B}	The Boltzmann constant
K_i	Force constant in ith umbrella sampling window
$k_{\theta,ijk}$	Force constant for angle between atoms i and j
$k_{l,ij}$	Force constant for bond between atoms i and j
LINCS	Linear Constraint Solver
L	Length of the edge in cubic simulation box

$L_{\max}$	Largest box matrix element
l_{ij}	Bond length between atoms i and j
$l_{ij,\,\mathrm{eq}}$	Equilibrium bond length between atoms i and j
$\mathbf{m}$	Reciprocal box vector in Ewald summation
m_i	Mass of particle i
MD	Molecular dynamics
N	Number of particles
N_w	Number of sampling windows in umbrella sampling
N_f	Number of degrees of freedom
n	Periodicity in the cosine series for dihedral potential
$\mathbf{n}$	Simulation box's index vector
$\mathcal{O}(N)$	Computational complexity as a function of particle number N
PCM	Pericellular matrix, also known as the glycocalyx
$P^{\mathrm{b}_i}(\xi)$	Biased probability distribution in ith umbrella sampling window
$P^{\mathrm{u}_i}(\xi)$	Unbiased probability distribution in ith umbrella sampling window
PME	Particle Mesh Ewald
PMF	Potential of mean force
p_i	A weight factor assigned to ith umbrella sampling distribution
$p(v_i)$	Probability as a function of velocity v for particle i
$\mathbf{P}$	Pressure matrix in Parrinello–Rahman pressure coupling
$\mathbf{P}_{\mathrm{ref}}$	Reference pressure matrix in Parrinello–Rahman pressure coupling
Q	Canonical partition function
q_i,q_j	Charges on atoms i and j
r_{cut}	Cut-off distance
r_{ij}	Distance between atoms i and j
$\mathbf{r}_i$	Position of particle i
$\mathbf{r}_{ij}$	Interparticle distance vector from particle i to j
T	Absolute temperature
t	Time
$t_{\max}$	Ending time of a simulation
Δt	Simulation timestep
V	Volume
VMD	Visual Molecular Dynamics, a visualization software
v_i	Speed of particle i
$\mathbf{v}_i$	Velocity of particle i
V_n	Fourier component in cosine series for dihedral potential
WHAM	Weighted Histogram Analysis Method
$\mathbf{W}$	Matrix Parameter in Parrinello–Rahman pressure coupling
W_w	Wiener process

$w_i(\xi)$	Harmonic biasing potential acting along reaction coordinate ξ
α_{ij}	Isothermal compressibility
β	Weight factor in Ewald summation
θ_{ijk}	Bond angle between atoms i, j, and k
$\theta_{ijk,\,\mathrm{eq}}$	Equilibrium bond angle between atoms i, j, and k
ϕ_{ijkl}	Dihedral angle involving atoms i, j, k, and l
$\phi_{ijkl,\,\mathrm{eq}}$	Equilibrium dihedral angle involving atoms i, j, k, and l
ϵ_{ij}	Energy scale in Lennard-Jones interactions
σ_{ij}	Length scale in Lennard-Jones interactions
ε_0	Permittivity of vacuum
ε_r	Relative permittivity of the material
τ_t	Time constant in temperature coupling
τ_p	Period of pressure fluctuations in Parrinello–Rahman pressure coupling
ξ	Reaction coordinate

1. INTRODUCTION TO THE WORLD OF SUGARS AND SIMULATIONS

Cells are the basic units of life. Every living organism is built from these tiny, diverse, and fragile objects that have perfected themselves through billions of years of evolution. Not long ago, it was still thought that cells would possess some mysterious force of life over the rest of the universe. Today it is, however, known that they are driven by the same laws of physics that, for example, make water to boil in high enough temperature or control the organization of electrons around a nucleus of an atom. This realization has not removed any of the thrill of studying these amazing biological constructs, but on the contrary, has made it even more exiting and rewarding. Indeed, understanding the cells and the various interactions they make is not only vital for perceiving the nature, but may profoundly increase the quality of our lives. When considering the cellular interactions in particular, it is not surprising that the interface between the cells and their immediate surroundings comprises the most interesting area of research.

On the surface of cells, on the border of two environments, lies a layer called glycocalyx [1–4]. It is a thin gel-like coat lining most of the cells from simple bacteria to higher eukaryotes like the animal cells. Glycocalyx consists of carbohydrates, proteins, and other biological components attached to a cell membrane either chemically by covalent bonds or physically by electrostatic interactions [4]. It forms a functional interface constantly sensing the environment, mediating cellular recognition and adhesion, and gating the molecules that try to enter or leave the cell, thereby rendering cellular interactions complex and well regulated events [5]. In this regard, the most important components of the glycocalyx are the adhesion molecules, cell surface proteins that control the cell-cell and cell-matrix interactions.

Due to its central role in cellular proliferation, one of the most interesting functional units of the glycocalyx is an adhesion molecule called CD44 [6]. It is a ubiquitous multistructural and multifunctional cell surface protein that, as its hallmark function, interacts non-covalently with a carbohydrate polymer called hyaluronic acid (HA) [7]. This adhesion interaction maintaining a vast HA network around the cell is also sensing the surrounding microenvironment and guiding the cellular functions through multiple biochemical signaling pathways [8]. As a result, CD44 has a vital role in cell survival, growth, and motility. Proper function of these funda-

mental processes is required for processes such as white blood cell homing, healing of injuries, embryonic development, and controlled cell death.

Given the central role of CD44-HA interaction in normal cellular homeostasis, and the vast abundance of both species, it is clear that their interactions must be regulated extensively [6]. Indeed, reflecting the dynamic nature of these interactions, some cells at a particular developmental stage need the HA binding ability for their normal functions, whereas other cells at a different phase of the development do not adhere to HA at all [9]. One very efficient way utilized by most cells to alter protein structures, and therefore their ligand binding properties, is to glycosylate them by covalently linking carbohydrates to their surfaces [10], or also by allowing changes in their conformations. Failure in these normal regulatory mechanisms, however, leads to severe pathologies and malignancies. Indeed, CD44 has been repeatedly linked to conditions such as prolonged inflammation and cancer propagation [11, 12]. As a matter of fact, CD44 has been confirmed to be a major player in the spread of tumor metastases, thus highlighting its central role in cellular function and rendering it as an intriguing subject for further research.

CD44 has been studied extensively for two decades. In addition to a high number of experiments concentrating on its pathophysiology, a range of studies have also focused on the molecular level details of its structure. The goal has been to fully understand the factors affecting the CD44-HA interplay. Overall, techniques such as mutagenesis assays [13, 14], spectroscopy measurements [15, 16], and computer simulations [17, 18] have been applied, yet many details of the binding process are still missing. Most importantly, the previous endeavors of deciphering this surprisingly complex chain of events have not accounted for the modular structure of CD44, with several glycosylation possibilities altering the binding process. Nor has anyone evaluated the details of how the receptor recognizes different HA sizes. Furthermore, due to somewhat controversial results between previous NMR spectroscopy [19] and x-ray crystallography studies [16], the biological relevance of the conformational changes in the structure of CD44 has remained unclear.

Solving the precise details of CD44-HA interplay could also have pharmacological implications. In other words, knowing how the regulation of the interaction is conducted in nature may facilitate the design and the targeting of novel drugs against malignancies such as cancer. Moreover, this information is of potential value in unraveling the function of other HA binding proteins [20] or more generic protein-carbohydrate interactions.

Admittedly, the study of these biomolecules is not a straightforward task. Experimental methods comprising the backbone of the research are often limited by the dynamic and delicate nature of soft matter systems such as the glycocalyx [21]. However, while the importance of experimental methods should not be underestimated,

they can be complemented with a variety of different computational methods. One particularly appealing technique for studying biological phenomena that occur at energy scales of weak physical interactions, like thermal fluctuations, is a method called molecular dynamics (MD) [22]. Despite having some inherent weaknesses of its own, MD is nowadays widely used to simulate the dynamics of biological many-body systems.

In this Thesis we employ atomistic MD to explore the role of glycosylation and HA chain length in CD44-HA interplay. We present several model systems, with three distinct CD44 glycosylation schemes, two CD44 conformations, and three HA sizes. Most importantly, we measure the effect of glycosylation to HA binding by calculating the thermodynamic binding free energies for the adsorption of HA oligomer to its binding site in CD44 with and without the added carbohydrates. Furthermore, comprehensive binding profiles and free energy profiles reveal how HA fragments of different size are recognized. Based on our simulation data, we also question the biological relevance of the conformational changes in stabilizing the carbohydrate-protein interaction.

Given the lack of knowledge about the glycocalyx and its components in general, our work is like solving a puzzle. However, instead of having all the ready-made pieces, we need to build them ourselves from scratch. The project may also have far-reaching implications, as it is an integral part of a larger attempt to model the whole glycocalyx-layer with atomistic precision.

This Thesis is divided into six chapters. The next chapter presents the biochemical background of glycosylation along with the detailed structure of the glycocalyx. The third chapter then gives a state-of-the-art review on CD44 and its functions with an emphasis on the interaction with HA. Next, the fourth chapter describes the computational methodology employed in this study. Namely, it explains the theory behind classical MD methods and related free energy calculations and lists the used model systems and simulation parameters in detail. The fifth chapter presents the simulation results and compares them to previous ones obtained from both experiments and computer simulations. The last chapter then reviews the main results, discusses their biological relevance, and concludes by contemplating the possible future directions.

2. HYALURONAN-RICH GLYCOCALYX ACTS AS A FUNCTIONAL INTERFACE OF CELLS

Glycocalyx is a fascinating and still somewhat mysterious layer lining various cell types. Due to its vast amount of molecular components it renders the cell-cell contacts totally different from those of simple liposome-liposome interactions [21]. The term "glycocalyx", referring to a sweet husk, was first used by Bennet et al. [23] who discovered in 1959 with electron microscopy that a polysaccharide coat was surrounding several vertebrate blood capillaries. Ever since, this intriguing and complex structure has been called with many names, such as "pericellular matrix" (PCM), "cell coat", or is simply referred to as the carbohydrate layer. Similarly, the definition of glycocalyx has various interpretations in the narrow yet constantly growing literature around the topic.

Although recently the endothelial glycocalyx [1, 3, 4, 21, 24] has evoked considerable interest due to its conspicuous appearance, other cells, too, express carbohydrates on their surface [2, 25, 26]. In this Thesis, the glycocalyx is given a broad definition referring to every carbohydrate-bearing material on the surface of bacteria and almost all animal cell types. In fact, the slime on the surface of a snail, fish, or intestinal epithelium can all be regarded as a glycocalyx. It should be noted, however, that this study focuses only on animal cells, and more precisely, on human cell types.

Furthermore, the distinction between the glycocalyx and the extracellular matrix (ECM) is not always properly defined. Indeed, despite sharing some integral components, such as the HA polymers, these two structures confer totally different functions. For instance, the ECM produced mainly by fibroblasts, chondrocytes, and osteoblasts, forms the mechanical platform to which other cells then adhere to. This concept is remarkably well illustrated in the decellularized organs where only the extracellular framework remains [27, 28]. In contrast, the glycocalyx, forming a functional interface between cells and the ECM, executes mainly dynamical adhesion and gating functions.

This chapter describes the biological milieu in which the CD44-hyaluronan interaction takes place in. The first section gives a brief introduction to basic biomolecules and their glycoconjugates, especially the ones important for the present study. The second section then presents the state-of-the-art knowledge on the composition of the

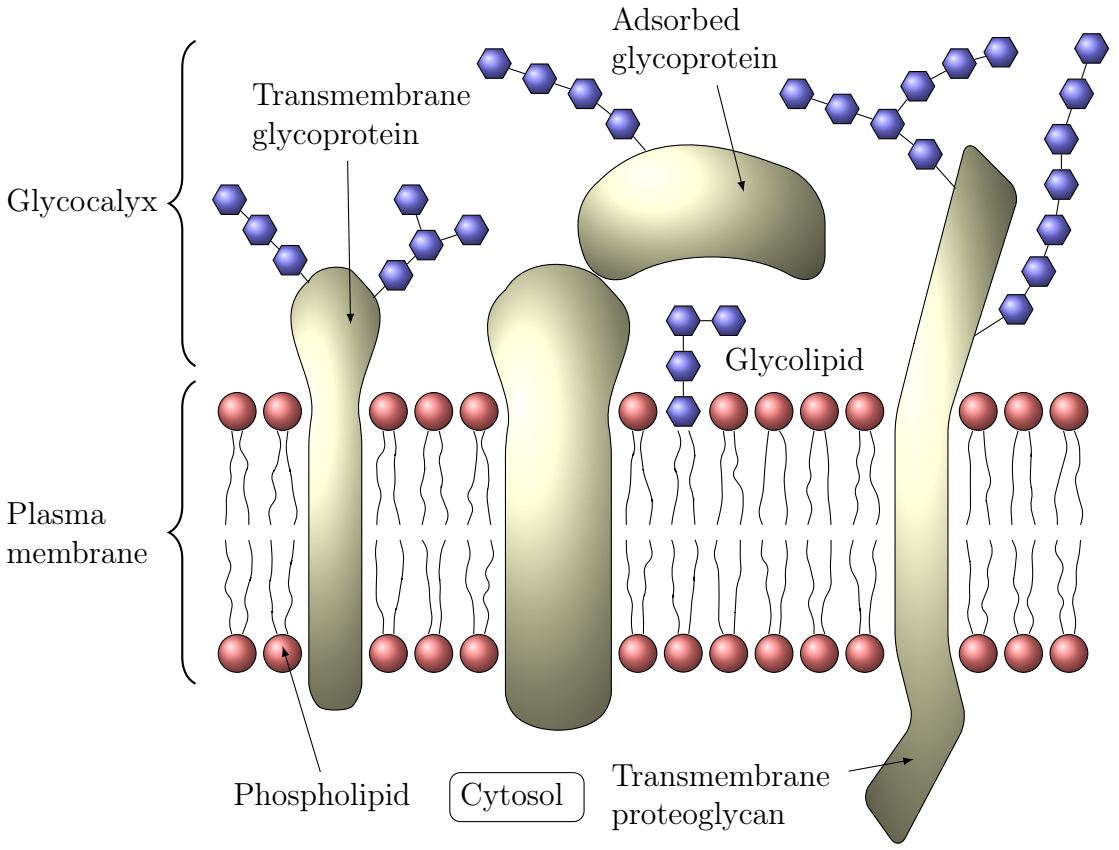


Figure 2.1: Collectively, glycocalyx and plasma membrane are formed from lipids, proteins, and sugars that combine to create several conjugate molecules.

glycocalyx. The last two sections briefly review the main properties and functions of HA, which is an integral part of our simulation models.

2.1 From Simple Biomolecules to Elaborate Glycoconjugates

Every cell expresses a distinct but vast number of biological macromolecules. They are the basic structural and functional units of our bodies and have been conserved through evolution due to their ability to carry out specific tasks in sustaining cellular functions. In practice, these molecules can be just small metabolites working as cofactors for larger enzymes or longer polymers with molecular weights over 500 kDa. Three basic classes of biomolecules −− proteins, lipids and polysaccharides −− coexist in the complex environment of the glycocalyx as illustrated in Figure 2.1 [4, 29, 30].

Proteins are long polymers of amino acids. From all the biomolecules they have the widest array of functions. Indeed, proteins act as catalysts, signal receptors, structural elements, and transporters. Lipids, on the other hand, are a broad class of amphipathic molecules that in a biological milieu serve e.g. as signaling molecules, structural components of membranes, storages of fuel, and pigments. Biological

lipids usually comprise a hydrophilic head group and two hydrophobic fatty acid chains connected to a certain backbone molecule. While cell membranes are primarily constructed of biological lipids, ordinary substances like waxes, vitamins, and fats can also be considered as lipids. Physically, membranes act as 2D fluid layers stuffed with transmembrane proteins having different compositions and compartments, such as cholesterol-rich lipid rafts.

Lastly, polysaccharides or carbohydrates are polymers of either mono- or disaccharides joined together by glycosidic bonds. These monomers include an array of simple sugars, such as glucose or galactose. In general, polysaccharides have two major functions inside the cells. First, like lipids they serve as energy-rich fuel stores releasing chemical energy, stored to their glycosidic bonds, when required. Second, polysaccharides are important structural elements in plants and bacteria, forming rigid cell walls around them, thereby providing them mechanical stability. Fundamental knowledge of all the aforementioned biomolecules is available in many excellent textbooks concerning proteins [26, 29, 31], lipids [26, 29, 32] and polysaccharides [26, 29, 33].

Glycosaminoglycans or simply GAGs encompass a special class of biopolysaccharides unique to animal cells. Unlike homopolysaccharides like starch or glycogen, they are linear disaccharide polymers that have been modified by sulphation or deacetylation of varying degrees [4]. In practice, disaccharides always comprise a uronic acid and a hexosamine, where the latter is either *N*-acetylglucosamine or *N*-acetylgalactosamine. Furthermore, the uronic acid is in most cases either D-glucuronic acid or L-iduronic acid [26]. The carboxylate group in uronic acid together with extensive sulphation leaves GAGs with a highly negative charge density. As a result, their conformation is largely dependent on sulphation patterns [34] and also allows specific recognition by carbohydrate-binding proteins, such as selectins. Furthermore, owing to the high negative charge density, GAGs usually assume an extended helix conformation in solution with carboxylate groups in alternate sides of the sugar chain.

In addition to roles as storage fuels and structural materials, polysaccharides act as information carriers. On the surface of a plasma membrane, they form the most prominent part of the glycocalyx that provides recognition sites for extracellular molecules and pathogens, thus enabling communication between cells and their surroundings. In comparison, inside a cell, sugars are employed in labelling proteins for transport, localization, or destruction. Indeed, almost always sugars, especially oligosaccharides (polymers with less than 20 monomers [33]), are covalently attached to a protein or a lipid to form so-called glycogonjugates [26]. These conjugate structures are named as glycoproteins and glycolipids unless the carbohydrate content exceeds that of the lipid or the protein, in which case they are referred

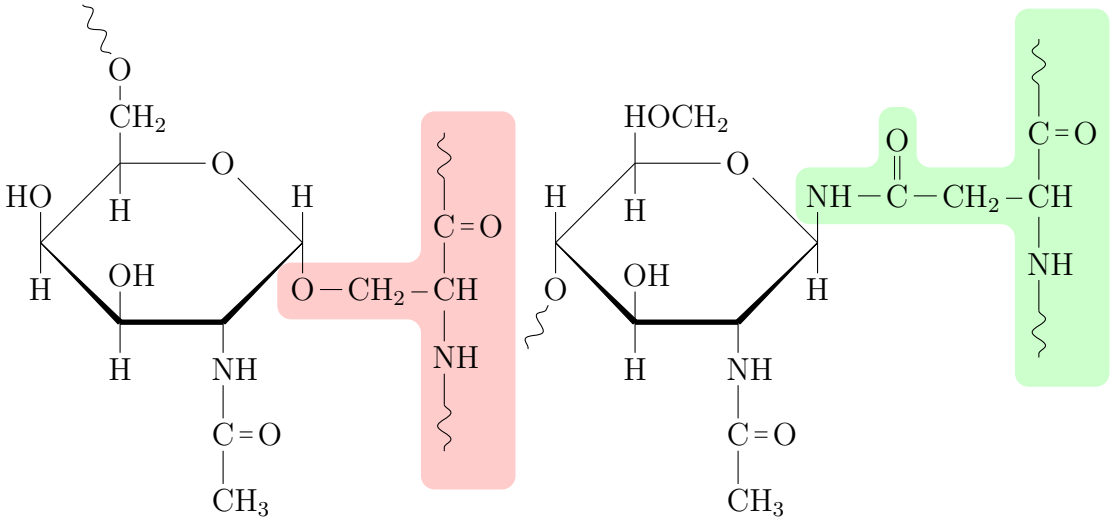
to as lipopolysaccharides and proteoglycans, respectively [33].

Proteoglycans are ubiquitous glycoconjugates of the PCM and ECM. They constitute one or more GAGs covalently linked to a protein that can be either attached to a membrane or secreted to the ECM. At first, proteoglycans assemble in the endoplasmic reticulum (ER) and in the Golgi complex, where a range of enzymes attach GAGs to serine (Ser) residues of the core proteins via a so-called tetrasaccharide bridge. Later, additional modifications in the Golgi will determine the final destination of the newly synthesized proteoglycans. One possible target destination is, for example, the glycocalyx where the proteoglycans bind extracellular ligands, thus activating them, or simply help to sustain an ideal local concentration of these soluble substances. Some secreted proteoglycans also form enormous supramolecular complexes called aggregates. These huge assemblies then contribute to the strength and resilience of many mammalian tissues [26].

Another important class of protein-sugar conjugates are glycoproteins. Like proteoglycans, they also have several carbohydrates of varying complexity attached to a core protein. The main difference between the two classes is, however, that glycoproteins have generally much shorter (2-15 monomers long), more branched, and more structurally diverse sugar residues attached via special glycosidic linkages [4]. Importantly, only a few carbohydrate species participate in these peptide links, namely *N*-acetylglucosamine (GlcNAc), *N*-acetylgalactosamine (GalNAc), galactose, mannose, xylose, and L-arabinose [33]. Glycoproteins are mainly found in the ECM, glycocalyx, and blood, although some cytosolic proteins can also be glycocylated. Demonstrating the vital role of these modifications, 50 % of all of the mammalian proteins are glycocylated and 1 % of all the genes coding for enzymes relate to glycocylation [26]. The attached oligosaccharides then shield the protein from cleavage and function as name tags for transport. They may also alter the polarity and stability of the protein that they are attached to.

The linking itself can happen in two ways as shown in Figure 2.2. First, an *O*-glycosidic link joins a carbohydrate to the hydroxyl group of a Ser or threonine (Thr) residue. Second, an *N*-glycosidic link attaches a carbohydrate to the amide nitrogen of an asparagine (Asn) residue. The latter linkage depends on a consensus sequence of three amino acids where the first is the linking Asn, the second is anything else but proline (Pro), and the third is either a Ser or Thr [26]. Still, not all of these sequences combine with carbohydrates, and on the other hand exceptions to the consensus sequence have also been reported [33].

The propensity of an Asn residue to be glycocylated depends on several factors [33]. On an atomistic level, the linkage is favorable when the hydroxyl group of Ser/Thr forms a hydrogen bond to the carbonyl group of an Asn side chain, decreasing the dissociation constant of the amide group, and thereby increasing



(a) *O*-linked (b) *N*-linked

Figure 2.2: (a) *O*-linked oligosaccharides form glycosidic bonds to the hydroxyl group of Ser/Thr side-chains (red). The illustration shows *N*-acetylgalactosamine at the reducing end of the sugar chain. (b) *N*-linked oligosaccharides connect to the amide nitrogen of an Asn residue (green) with *N*-acetylglucosamine as the terminal sugar.

its affinity for the carbohydrate. Other elements influencing the formation of an *N*-glycosidic linkage are dictated by the surrounding environment. For example, enzymes performing the concatenation are highly tissue specific. Furthermore, besides the primary structure, also the tertiary structure around the consensus sequence mediates the affinity for the carbohydrates. Lastly, but rather interestingly, the tendency to be glycocylated seems to be reduced the closer the sequence is to the C-terminal end of the protein.

The most common type of attached carbohydrate in mammalian glycoproteins is the so called "core pentasaccharide" portrayed in Figure 2.3. Usually these kinds of oligosaccharides are highly specific for a certain individual, like the carbohydrates defining our blood groups. However, the core pentasaccharide is neutral in that respect, and thereby induces no immune response. On a structural level, the reducing end of the pentasaccharide Man-α1 $\rightarrow$ 6(Man-α1 $\rightarrow$ 3)-Man-β1 $\rightarrow$ 4-GlcNAc-β1 $\rightarrow$ 4-GlcNAc-β1 $\rightarrow$ or simply $Man_3GlcNAc_2$ attaches to the side chain amide nitrogen of the Asn residue through an *N*-glycosidic linkage. Usually, the distal mannoses are extended by the addition of other monosaccharides, such as sialic acids, to their ends [33].

Just like proteins, also lipids can be conjugated with carbohydrates [26]. It is usually the membrane sphingolipids whose head groups are replaced with oligosaccharides like sialic acids. These glycolipids play an important role, for example, in signal transduction. Consequently, the brains and neurons are often rich in these conjugate lipids.

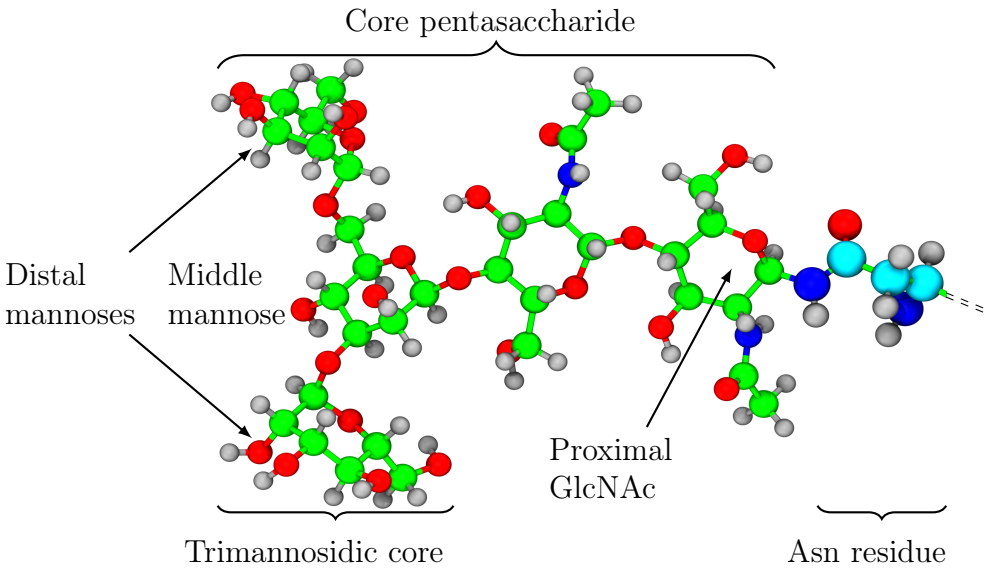


Figure 2.3: Core pentasaccharide attached to an Asn residue of a protein and represented with a ball-and-stick model using Visual Molecular Dynamics (VMD) software [35] and Tachyon ray tracing library to render the image. Basic structure and common naming conventions are illustrated with braces and arrows. Hydrogen atoms are colored gray, oxygen atoms are red, and carbon atoms that belong to the sugar are green. For the sake of clarity, protein carbons are colored cyan.

2.2 The Glycocalyx Houses a Wide Array of Biomolecules

Glycocalyx is an intricate, highly complex self assembling 3D network. In addition to numerous polysaccharides and glycoconjugates, it also includes a large variety of adsorbed components with no clear boundaries between those and the synthesized ones [4]. Furthermore, in normal physiological conditions, the glycocalyx is subjected to high turnover rates and thus experiences a dynamical balance between carbohydrate synthesis and shedding-dependent alterations [1]. Due to this dynamic and soft nature, the glycocalyx is extremely hard to define geometrically [21, 36]. In spite of this, many of the individual components are nowadays recognized and characterized [4]. Their distribution and intermolecular interactions have, however, remained largely unidentified, and if such data exists, it is only qualitative and often indirect. Figure 2.4 depicts the complex structure of the glycocalyx with only a few of the main components presented explicitly.

Overall, the glycocalyx contains five different GAG species, namely hyaluronic acid (HA), chondroitin sulphate (CS), heparan sulphate (HS), dermatan sulphate (DS), and keratan sulphate (KS) [4]. While HA is the most abundant GAG in the glycocalyx forming its structural basis, the others are also highly important in modulating and cross-linking the pericellular environment.

The glycocalyx connects to a plasma membrane via so-called backbone molecules,

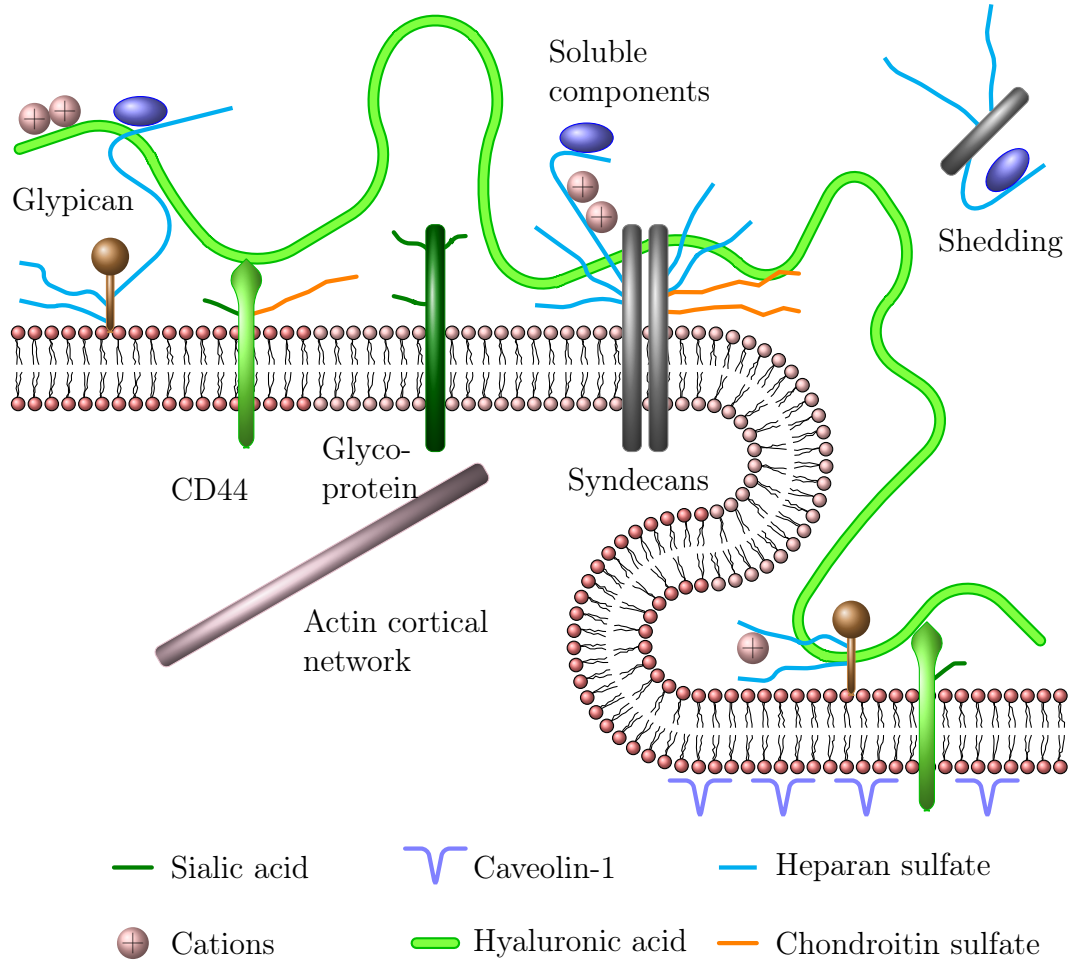


Figure 2.4: Schematic figure representing the main components of the glycocalyx. Reproduced from ref. [37].

namely glycoproteins and proteoglycans. Generally, glycoproteins reside in special cholesterol-rich lipid rafts, to which the caveolin-1 proteins also attach from the cytosolic side of a plasma membrane in order to bend it and to form cave-like structures called caveoli [3]. For example, glycocylated adhesion proteins like integrins, immunoglobulins, selectins, and isoforms of CD44 often concentrate on these rafts. In addition, glycoproteins along with some proteoglycans connect the glycocalyx to the cytoskeleton, and thereby enable the cell to sense the surrounding environment and adapt mechanically to each change [3, 5].

Proteoglycans of the glycocalyx vary largely in their size, number of attached GAGs, and their interactions with a plasma membrane. For instance, glypicans [38] are anchored with a glycosylphosphatidylinositol (GPI) anchor, and syndecans [39] are transmembrane proteins, while perlecan [40], mimecan, and biglycan [41] are secreted. Interestingly, the GAG-ratios of these proteoglycans have been observed to change not only according to intracellular, but also extracellular stimuli [42].

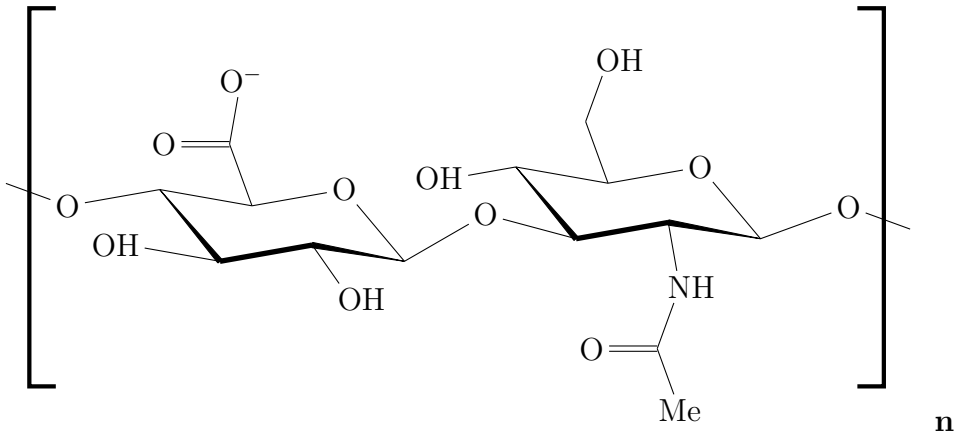


Figure 2.5: The chemical structure of hyaluronic acid.

This highlights the dynamic nature of the glycocalyx constantly subjected to various alternations and stimuli.

Despite their non-covalent attachment, the adsorbed components are an important ingredient of the glycocalyx. Especially in the vascular endothelium, many blood-borne proteins and soluble proteoglycans, along with cations of all sorts, adhere to the surface of the cells [3, 4]. As a result, they sustain the charge-selectivity within the thick network of negatively charged sugars. Moreover, the soluble components are suspected to function as cross-linkers for GAGs, such as the HA.

2.3 Hyaluronan is the Primary Component of the Glycocalyx

Hyaluronan or hyaluronic acid is a linear, multifunctional, and highly anionic GAG forming the basis of both glycocalyx and ECM [2, 7, 25, 43]. It contains alternating *N*-acetyl-D-glucosamine (GlcNAc) and D-glucuronic acid (GlcUA) residues linked to each other via a $\beta(1 \rightarrow 4)$-glycosidic linkage, while the dimers attach to each other through $\beta(1 \rightarrow 3)$ bonds. HA is also extremely abundant in most tissues of the human body. For instance, synovial fluid and vitreous humour contain high concentrations of this polymer, because it acts as a lubricant and gives these fluids their jelly-like consistency. Furthermore, the name hyaluronic acid itself was derived from this glassy and translucent appearance, as the Greek word "hyalos" refers to glass [26]. Figure 2.5 presents the chemical structure of one HA dimer unit.

HA was first purified in 1934 and its chemical structure was solved twenty years later [25]. At first, it was thought to serve as merely a lubricant and structural scaffold, yet later HA has been shown to play a role at least in adhesion, migration, and proliferation of various cell types, along with water homeostasis, filtering, exclusion, scavenging of free radicals, and mediating the angiogenesis in most of the mammalian tissues [44]. Moreover, currently HA is widely employed in the cosmetic industry [45] and has also been recently utilized in various biomaterial applications

serving as tissue engineering scaffolds [46] or drug delivery devices [47].

Compared to other GAGs, HA shows some unique physical properties. For example, it is regarded as the longest molecule of the human body with molecular weights up to 6-7 $\times 10^6$ Da [48]. This weight covers a remarkable 2000 to 2500 disaccharide units and spans polymer lengths of 2-25 μm [7]. The incredible magnitude combined with the highly negative charge density originating from the carbonyl groups in uronic acids gives HA its unique rheological properties, such as dynamic viscosity and elasticity varying with concentration, molecular weight, and experienced shear forces. Furthermore, the high flexibility of the HA polymer renders it an excellent space-filling molecule, cramming voids in tissues or even acting like a molecular spring.

In addition to being much longer and having different chemical structure, HA differs from other GAGs in the way it is synthesized. That is, unlike other GAGs, HA is assembled on the cytosolic side of a plasma membrane by proteins called HA synthases (HAS). Instead of being modified afterwards, HA chain is extruded straight to the extracellular space or retained at the glycocalyx. However, distinct modification patterns arise from polymers of various lengths as they have been shown to induce different physiological responses [49]. Indeed, HA is synthesized and released in pieces of over 2×10^6 Da [50, 51], while the smaller fragments with less than 20 carbohydrate monomers result from trauma or, by contrast, form through hyaluronidase enzymes cleaving the longer HA polymers [52]. Interestingly, high molecular weight polymers of HA seem to inhibit the formation of capillaries, whereas the much shorter fragments induce it [49]. Another study showing only the short fragments to inhibit the formation of glycocalyx in human aortic smooth muscle cells *in vitro* [53] further highlights the discrepancy between the different-sized HA polymers.

Still, one additional aspect separates HA from other GAGs – the binding to proteins. While GAGs usually link to proteins covalently, HA is bound non-covalently by a range of receptors commonly referred to as hyaladherins. They almost always contain the HA-binding Link module, which is a disulphide-linked domain of around 100 amino acids acting as the hallmark of a protein family called the Link module superfamily [25]. This name refers to a cartilage link protein involved in the formation of aggregate [54] structures, in which the HA chain forms the basis to which proteoglycans called aggrecans attach via the link protein in a brush-like fashion. Other members of this superfamily include laylin [55], LYVE-1 [56], and stabilin-1 [57]. Yet, the primary receptor for HA is the cell-surface protein CD44 [58]. In addition to these, HA is bound by proteins lacking the Link module, namely by RHAMM (Receptor for Hyaluronan-Mediated Motility) [59], which despite having no homology to other HA-binders, is able to compensate the function of CD44 in

Table 2.1: A list of common hyaladherins and their locations inside human tissues.

Hyaladherin	Description	Ref.
CD44	Glycocalyx	[58]
RHAMM	Glycocalyx	[59]
LYVE-1	Glycocalyx	[56]
Cartilage link protein	ECM	[60]
Aggrecan	ECM	[54]
Versican	ECM	[53]
Neurocan	ECM	[61]
Brevican	ECM	[62]
TSG-6	ECM	[63]
cdc37	Intracellular	[64]
P32	Intracellular	[65]

CD44-knockout mice. Lastly, HA polymers also sometimes retain at the cell surface through their interactions with the HA synthases. In conclusion, many elements that affect HA metabolism have been identified, yet similarly to the glycocalyx, the molecular level details are still largely unknown. Table 2.1 summarises the main HA-binding proteins.

Depending on the surrounding environment and the available stimuli, HA may assume various configurations, as illustrated in Figure 2.6. For example, in the ECM it forms highly cross-linked meshworks interacting with other ECM components, such as collagen. These meshworks are stabilized by HA cross-linkers like TSG-6 (35 kDa secreted product of the tumor necrosis factor stimulated gene-6) [63, 66], yet another protein with the Link module. In contrast to cross-linked networks, cells also produce long cable-like HA structures [67–69] that seem to appear only under specific stimuli, such as stress in the endoplasmic reticulum, viral infection, or inflammation.

2.4 Hyaluronan Dictates the Functions of the Glycocalyx

The function of the glycocalyx is largely governed by the behaviour of HA retained at the cell surface. In practice, these functions include acting as a permeability barrier, facilitating leukocyte rolling upon inflammation, scavenging free radicals, and transducing mechanical cues from the environment to cellular signals. Traditional ways to study this HA-rich layer entail methods such as atomic force microscopy [70], electron microscopy [71], optical traps [72], and particle exclusion assays [2]. The latter refers to a procedure where a suspension of fixed erythrocytes is allowed to settle, and subsequently a clear zone surrounding the sample cells is made visible because of exclusion of the red blood cells by the HA coat. As a hyaluronidase

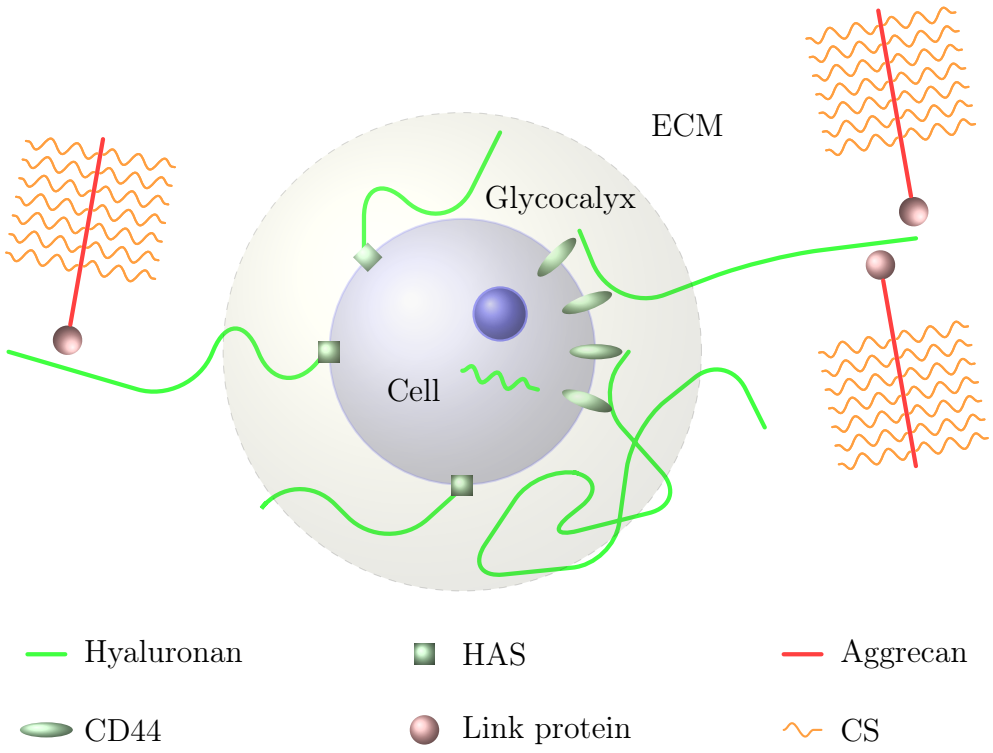


Figure 2.6: Schematic illustration of intra-, peri-, and extracellular hyaluronan surrounding a cell like a chondrocyte.

treatment removes the optically clear zone, the integrity of glycocalyx is indeed confirmed to be HA-dependent.

Owing to the extremely high density of HA meshworks, the glycocalyx is an excellent permeability barrier, gating all the molecules trying to enter or leave the cell. Behaviour like this is especially important in the vascular endothelium where the absorption needs to be strictly controlled [73]. Indeed, blood hematocrit, the volume percentage of red blood cells in the blood, experiences a two-fold rise when capillaries are treated with hyaluronidase [74], thereby indicating the presence of a thick glycocalyx. Consequently, the endothelial permeability for high molecular weight molecules is far less than that for smaller ones [75]. Besides steric hindrance, also charge plays a role when molecules try to pass through the cell coat. For instance, neutralization of the glycocalyx with salt solutions in rat mesenteric small arteries increases the intake of 50 kDa dextrans by a factor of two [76]. Due to its charge and the osmotic forces, the glycocalyx is also an important regulator of water homeostasis, creating an effective barrier for water and thereby guiding fluid flows in tissues and protecting them from edema [3]. A recent study in fact shows that a fixed polysaccharide network orders water and offers marked resistance to the bulk flow of solvent [77]. In the light of this, it is no surprise that treating capillaries with hyaluronidase increases the blood hematocrit so significantly.

In addition to repulsion, the glycocalyx exhibits adhesive properties as well. It has been implicated to play a central role, for instance, in leukocyte rolling. This refers to a process related to inflammation where white blood cells, also known as leukocytes, roll on the surface of the endothelium, thus creating right circumstances to enter the underlying tissue [78, 79]. This blood-to-tissue transition is, in turn, called extravasation. Traditionally, adhesiveness is correlated to integrin molecules at the surface of the leukocytes and the endothelium [80, 81]. The role of HA-rich glycocalyx has only recently begun to be unravelled [4], and interestingly, it seems to possess a dual capacity in the recruitment of white blood cells.

On one hand, the glycocalyx promotes adhesion. Being a rather loose structure it may allow the membrane extrusions of the leukocytes, where the adhesion molecules usually reside, to reach the plasma membrane relatively easily. Furthermore, the glycocalyx is known to mediate the early long-range interactions with the surfaces of tissue culture substrates, before a firm integrin-mediated adhesion commences [82]. Similarly, the cable-shaped HA structures are clearly pro-adhesive, promoting adhesion of white blood cells both *in vitro* and *in vivo* [68, 83]. Indeed, they seem to possess unique leukocyte-binding properties and may potentially influence the clustering of HA-receptors on the surface of white blood cells [67].

On the other hand, the glycocalyx dampens the leukocyte adhesion. Steric hindrance seems to play an integral role since most adhesion molecules are much shorter than the thickness of the glycocalyx [84, 85]. Furthermore, HA is though to be an antiadhesive lubricant promoting cell rounding and detachment. Supporting this hypothesis, a decreased synthesis of HA due to an HAS antisense RNA is found to reduce the migration of keratinocytes [86]. Further supporting the anti-adhesive character is the fact that highly flattened and spread cells, like the smooth muscle cells, produce very little or no glycocalyx [2]. When both the cell and the substrate express HA, no attachment takes place, but when only other one has it, the connection is established [2]. Overall, adhesiveness seems to depend on reasons such as local concentration, cross-linking, synthesis rate, and degradation of HA, while at the same time being regulated by factors like size, amount, cellular location, and malleability of this long polymer.

Yet another matter affecting leukocyte rolling and attachment is the shedding of the glycocalyx. In other words, certain chemical and mechanical cues are observed to shed the HA chains of the glycocalyx, so that the synthesis of GAGs and the shedding exist in a dynamic balance [78]. This behaviour is also believed to be an important mediator in inflammatory response [87]. In practice, free radicals and various chemoattractants released by inflammation would activate matrix metalloproteinases, zinc-containing proteolytic enzymes that degradate components of the ECM. These enzymes would in turn facilitate the shedding of the glycocalyx. Sub-

sequently, leukocytes in blood would get an easy access to the plasma membrane of the endothelium [1]. This could, as such, imply that the glycocalyx acts as a barrier for white blood cells in uninflamed tissues, denying their access to healthy parts of the body. However, inflammation is not the only stimulus to which the cell coat is known to react to. Shedding can, indeed, be facilitated by a range of factors such as hyperglycemia [88], septic shock [89], presence of oxidized LDL [90], ischemia [91], high blood shear stress [92], many biochemical agents [93], and free radicals [94].

3. CD44 IS THE PRIMARY RECEPTOR FOR HYALURONIC ACID

CD44 is a highly abundant molecule in our bodies. It is present in most cell types, but is especially plentiful in leukocytes, fibroblasts, endothelial cells, white matter of the brain, and smooth muscle cells [6]. In addition to these normal cell types, CD44 is encountered in a range of cancers [11, 12, 95–97]. It was first described as brain-granulocyte-T-lymphocyte antigen [98]. After its discovery, at the early 1980s, it has been referred to with several structural and functional names, such as $GP90^{Hermes}$, ECM receptor II, hyaluronate receptor, phagocyte glycoprotein-1, homing cell adhesion molecule, glycoprotein i5, Ly24, and HUTCH-1. However, when it was cloned for the first time at the late 1980s, the name CD44 was established [6]. A following sequence comparison revealed over 50 % homology with the cartilage link protein, thereby suggesting that CD44 might also be an HA binding protein. In 1991, a rat pancreatic tumor cell line was induced to metastasize by the transfection of complementary DNA (cDNA) encoding a specific CD44 isoform [99], and after this surprising discovery CD44 has been an interesting target for a range of pathological and biochemical studies that have thus far implicated it to play a crucial role in both cancer progression and cellular motility.

This chapter presents the latest knowledge on CD44 structure and functions, with a focus on its interactions with HA. The chapter is divided into four sections that each describe a different part of the CD44 biology. The first section gives a solid overview of the CD44 structure and its genomic organization. The second chapter then presents the main physiological implications of the CD44-HA interactions, thus helping the reader to understand the biological importance of this molecular-level event. Next, the third section summarizes a complex, yet important, aspect of native CD44 function, the regulation of HA binding. The chapter then concludes by a review of the relevant structural and computational studies previously conducted on CD44.

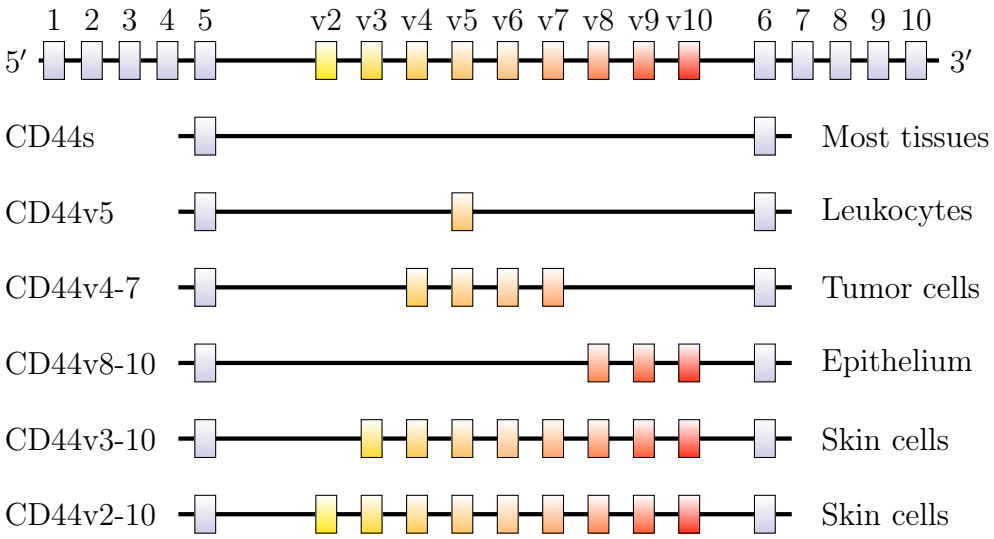


Figure 3.1: Exon map of the CD44 genome. The left column gives the names of the corresponding isoforms, while the right column specifies the cell or tissue where it is normally found.

3.1 CD44 is a Membrane Protein with Functionally Distinct Domains

CD44 is a transmembrane glycoprotein. It is coded by a single gene that is, however, subjected to extensive alternative splicing, enabling the cell to generate multiple unique isoforms of the protein. In humans this gene is located at the short arm of chromosome 11 [100] and covers roughly 50 kb of genomic data [101]. In mice it contains 20 exons [102] from which the first five are constant and thus, termed as the 5' constant region. By contrast, the next 10 exons are subjected to alternative splicing and thereby create a so-called variable region, designated as exons v1–v10 or simply as exons 6–15 in the sequential scale. The last five exons in the gene are called the 3' constant region, while the final exons, 19 and 20, are again alternatively spliced [102]. Importantly, human CD44 has only 19 exons [101] because the v1 tract does not exist. The naming of the gene, however, still follows the same convention as with the mouse gene, so that in humans v2 is the first variably spliced exon (Figure 3.1). It should also be noted that this study focuses only on the human CD44, albeit it has 88 % sequence homology with the mouse protein.

Given the large variety of possible exon combinations, CD44 occurs in various isoforms [103–105]. The most common of these is the standard CD44, abbreviated as CD44S or CD44H because it is especially abundant in hematopoietic cells [106]. This standard form has no variable exons, so the 5' region is directly attached to the 3' region. After the cleavage of the N-terminal 20 amino acid long signal peptide, CD44S translates into a 341 amino acid long protein with a predicted

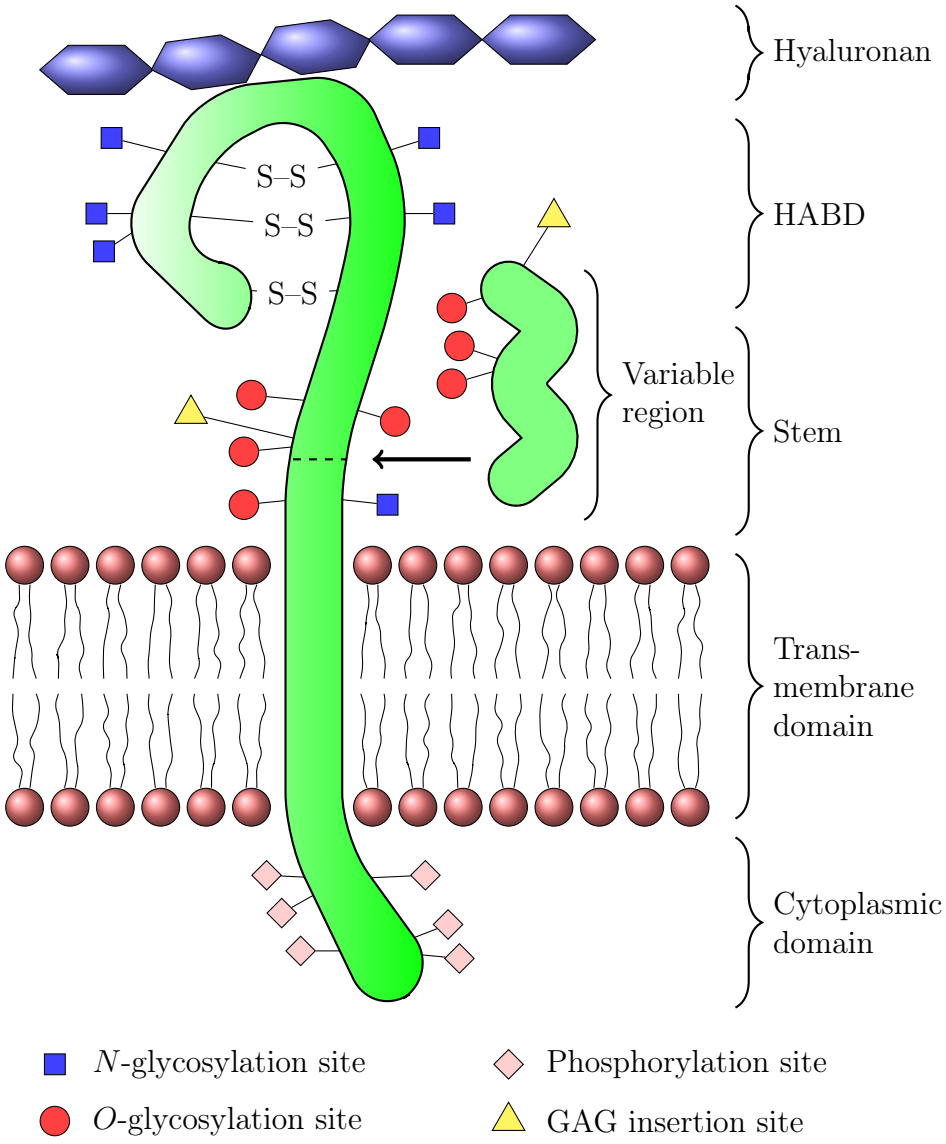


Figure 3.2: Schematic figure representing the different domains and posttranslational modification sites of a membrane-bound CD44.

molecular weight of 31–38 kDa [6]. The observed molecular weights are, however, around 85–95 kDa. This apparent discrepancy can be explained by the insertion of extensive posttranslational modifications like glycosylation, GAG attachment, and phosphorylation from which the last takes place only in the cytoplasmic parts of the protein, while the first two modify the extracellular part of the protein [6].

CD44 structure can be divided into three main parts: extracellular region, transmembrane segment, and cytoplasmic tail. The first is further divided into an N-terminal HA binding domain (HABD) and to a stalk-like stem region that connects the HABD to the transmembrane segment [15]. A schematic presentation of the CD44 structure illustrating these structural principles is shown in Figure 3.2.

The 158-residue-long globular HABD [15] is encoded by the exons of the 5' con-

stant region [6]. It contains six cysteine residues stabilizing the globular fold by creating three intraprotein disulphide bridges [107]. As the name implies, the HABD comprising the Link module (residues 32 to 123) enables the binding of the HA polymers [15, 16, 108]. Importantly, HABD retains its ability to bind HA even when expressed alone, without the rest of the protein [108]. This is a feature that has made it possible to study the details of CD44 structure and function, even though the whole transmembrane protein has been difficult to clone and crystallize.

The stem part is coded by the first two exons of the 3' constant region. Apart from linking the HABD to the transmembrane segment, very little is known of its structure. It is, however, known to contain multiple O-glycosylation and GAG insertion sequences, and therefore to be heavily glycosylated in its native state [10]. The stem also contains sites for proteolytic cleavage [109, 110], a process that might be vital for cell motility, as it can rapidly terminate the adhesion interactions. Most importantly though, the stem is the structure to which the alternatively spliced parts are inserted in variant isoforms [111]. In principle, over 800 different membrane-bound CD44 isoforms could be generated [6, 8], but because the trancription process is highly regulated by mitogenic signals not all of them are encountered in real life cell populations. Notably, the variant exons introduce additional glycosylation sites to the stem of the protein. Along with ordinary N and O-glycosylation sites some of the variant parts, like v3 and v6 exon products, contain specific Ser-Gly-Ser-Gly motifs for the incorporation of HS and other GAGs [107, 112–115].

After the extracellular parts, the next 21 amino acids (residues 269-290) coded by exon 18 constitute the short transmembrane segment. It contains one cysteine residue having the ability to form disulphide bridges, and it is therefore believed to be involved in CD44 oligomerisation [116], yet another factor regulating CD44 function [6]. The transmembrane segment is also thought to be responsible for locating CD44 into specific lipid rafts called glycolipid enriched microdomains (GEMs) [117]. This GEM-association varies according to cell type, but is regarded as a crucial aspect in CD44-mediated signal transduction, as GEMs harbor most of the signaling molecules that associate with the cytoplasmic tail of CD44.

The cytoplasmic tail region can be coded by either exon 19 or 20. As a result, the length of the tail is either 3 or 72 amino acids, respectively [6]. The long tail is, however, more common and it contains binding sites for many intracellular adaptor molecules along with six conserved phosphorylation sites [118]. This is also the part considered to be responsible for the subcellular localization, and mediating the clustering of CD44 through interactions with the cytoskeleton [6].

3.2 Ligand Binding Induces CD44-Mediated Signaling and Adhesion

CD44 functions can be divided into at least three specific categories: a platform for enzymes and substrates, a co-receptor for signal-transducing molecules, and a simple structural adhesion molecule [8]. Unlike the others, the first function is not strictly dependent on HA binding. That is, when residing in the glycocalyx CD44 acts as a platform to growth factors and enzymes, thereby increasing the capacity of these molecules to interact with their own ligands, and thus lowering the threshold for cellular signal tranduction [119]. In addition to HA, CD44 has been known to have also other, secondary ligands like osteopontin and mucosal vascular adressin [120–122]. The biological relevance of these interactions is not clear, but these molecules seem to have binding sites especially in the variant region and have also been implicated in cancer or prolonged inflammation [123].

The second main function means, in essence, that the CD44-HA binding activates several intracellular signaling pathways leading to changes in cell motility, shape, and survival [6, 8]. Instead of possessing an intrinsic catalytic activity, the CD44 tail region affiliates with several adaptor molecules with the ability to initiate migration and cell survival-related signaling pathways [124–128]. For instance, Src family of receptor tyrosine kinases (RTKs) is constitutively associated with CD44, activating a Ras signaling pathway trough protein kinase C (PKC) [129], thus upregulating genes that promote cellular migration and invasiveness [130]. Further fortifying the role of CD44 in cancer progression, PKC also upregulates downstream factors inhibiting apoptosis, the process of programmed cell death, and altering the body's drug resistance [12, 131]. In fact, most of CD44 adaptor molecules seem to be related to cancer progression by controlling these vital cellular processes.

A group of transmembrane RTKs called the ERBB protein family is often expressed in various tumor cell types. Interestingly, at least two studies have demonstrated CD44 to form heterodimers with these proteins, especially with ERBB2, and that this interplay is highly dependent on CD44's interaction with HA polymers [127, 132]. The CD44-ERBB2 complex indirectly activates the Ras signaling cascade [127] through a chain of secondary adaptor molecules associating with it on the cytoplasmic side of the a membrane, as shown in Figure 3.3.

Band 4.1 superfamily proteins, such as one called ERM [133], are yet another class of CD44 partner proteins utilizing the common Ras pathway. In addition to activating the growth permitting state in a cell cycle, the active ERM enhances the incorporation of variant exons to CD44s being assembled at the ER [134]. However, there is normally a delicate balance between ERM and its antagonist called Merlin, which acts as a competitive inhibitor for ERM when binding the cytoplasmic tail of

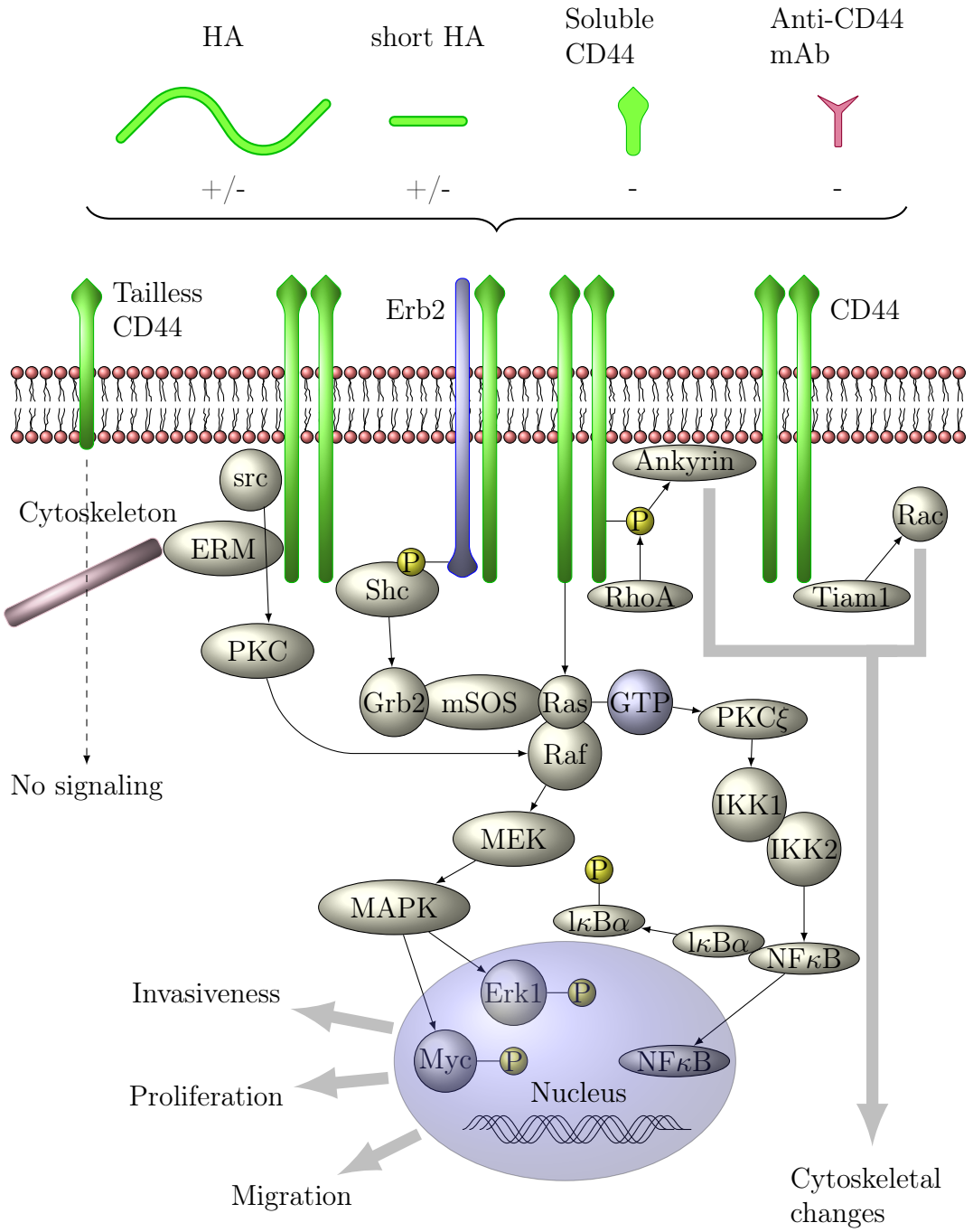


Figure 3.3: CD44 induced signaling (reproduced from Ref. [25]). The signals are transmitted through accessory molecules that activate several signalling cascades that in turn active nuclear transcription factors or other effector molecules. Most of the observable down-stream effects relate to migration and cell shape changes. Anti-CD44 mAbs and soluble CD44 HABDs can inhibit the signaling, whereas the size of HA can have either positive or negative effect.

CD44 [135]. Which one of the two is more powerful depends on the ruling cellular conditions. For instance, mitogenic treatment seems to favor growth, while a high cell density together with the presence of high molecular weight HA inhibits growth by activating Merlin [135]. This behavior is at the same time an intriguing example of the effect of HA size on CD44-mediated cell signaling.

Some CD44 adaptor molecules, including ERM, also directly cross-link CD44 to the cytoskeleton. These interactions then lead to rapid reorganizations of the actin cytoskeleton causing the formation of membrane protrusions, such as lamellipodia [128]. For example, T-lymphoma invasion and metastasis inducing protein 1 (Tiam1) promotes cytoskeletal reorganization through Rac1 GTPase enzyme [125] and a protein called Ankyrin links CD44 to cytoskeleton after activation by a CD44-bound Rho kinase [136, 137]. Interestingly, Ankyrin only accumulates underneath CD44 upon its binding to HA [138], again displaying the vital role of HA in the CD44-mediated functions.

Remarkably, the CD44-mediated signaling seems to be totally dependent on the interaction with HA. Whether this is due to accumulation of CD44 or some other reason, it can be readily demonstrated by disrupting the constitutive CD44-HA interactions using three different methods. First, the utilization of exogenic HA oligomers has been effective in displacing the polymeric high-affinity HA [139]. Second, soluble CD44 HABDs competitively bind to HA, and thereby hinder the native CD44-HA interactions [140, 141]. Third, CD44-specific siRNAs or mAbs are able to block the binding of HA by disrupting the translation of the protein or by competing for the same binding sites, respectively [6]. Applying these methods will result in an attenuation of the CD44-associated cell signaling [95, 142] and, for example, dissociation of the complex between CD44 and ERBB2 [127]. Although, it must be noticed that the size of HA may also play a significant role in determining the level, and even the malignancy, of the CD44-dependent signaling. In one occasion, for instance, shorter fragments ($\leq$ 267 kDa) were able to activate the nuclear transcription factor NF$\kappa\beta$ in several human cell lines, while the longer, high molecular weight, polymers had an inhibitory effect on the same process [143].

In addition to platform and co-receptor roles, CD44 can be considered simply as a ligand-binding protein facilitating cell traffic through reversible interactions. Indeed, cellular migration is most prominent in HA-rich spaces, such as in wounds undergoing repair, in inflamed tissues, or in close proximity to tumor cell mass [8, 12]. The dependency of these processes on the CD44-HA interactions may again be evidenced by blocking the cell movement with CD44-specific mAbs, soluble HA, or treatment with hyaluronidase [144]. Furthermore, the interplay between CD44 and HA ultimately gives leukocytes the ability to roll on endothelial surface in a velcro-like fashion, creating suitable circumstances for successful extravasation [145]

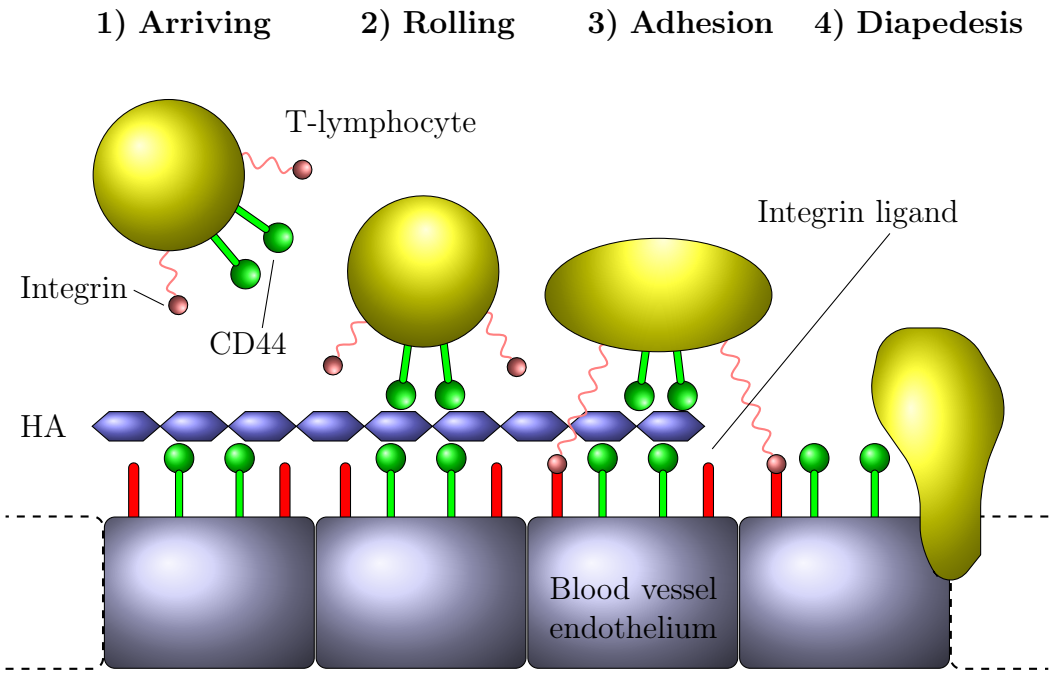


Figure 3.4: CD44 in leukocyte extravasation. Leukocytes in the blood stream adhere to the endothelium surface where the CD44-HA interplay takes care of the initial rolling interactions before strong integrin-mediated adhesion.

(Figure 3.4). Not surprisingly, it is also the same mechanism that is exploited by tumor cells metastasizing to secondary sites through blood or lymph circulation [12].

3.3 CD44 is Subjected to Multiple Levels of Regulation

Every CD44 expressing cells does not bind HA. This status may, however, alter very rapidly in response to changes in the surrounding conditions or intracellular stimuli. The dynamic switching of binding affinity is undoubtedly crucial for cells migrating through HA-rich matrices and for the steps of HA-dependent leukocyte rolling. Yet, no single mechanism controlling this behavior can be pointed out [8]. Instead, given the importance of the CD44-HA interaction, several levels of regulation are probably needed for the normal function. The binding of HA to CD44 is, indeed, collectively affected by at least factors like glycosylation [10, 16, 146], expression of variant exons [147, 148], conformational changes [16–19], cleavage of the cell surface CD44s [109, 110], clustering of CD44s [149], differential orientation of the CD44s at the cell surface [149], internalization of the HA polymers [150], and variations in the HA chain length [14, 149]. However, the relative contribution of these non-mutually exclusive mechanisms is currently unknown. Figure 3.5 summarizes the different factors having the ability to regulate the CD44-HA interactions.

One of the simplest way to regulate HA-binding is to recognize various sizes of HA fragments. CD44 has been shown to coordinate a minimum of six HA residues,

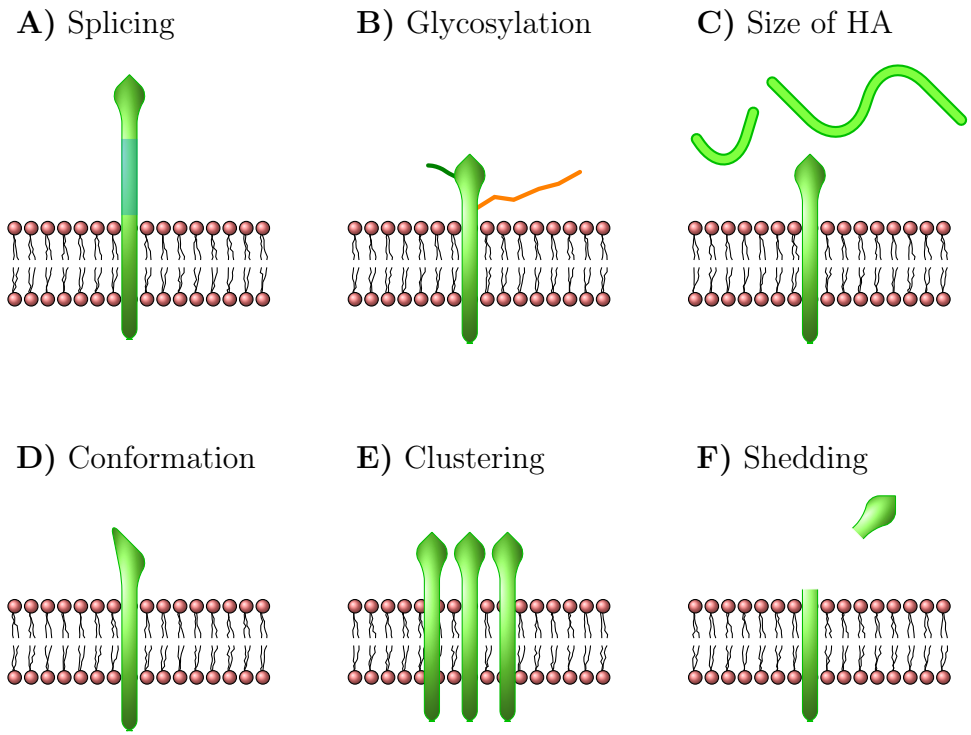


Figure 3.5: Factors that can regulate CD44-HA interactions.

or three disaccharide repeats, yet only after eight carbohydrate residues HA covers the full length of the binding site [14]. Whether these differences in binding affinity bear any physiological relevance is currently unknown. However, the evidence shows at least that low molecular weight HA (20–500 kDa) promotes cell cycle progression, while higher molecular weight fragments ($\geq 10^6$ kDa) inhibit the same process, rendering the cells quiescent and adhesive [44, 151]. Similarly, the shorter HA fragments induce inflammation, whereas the longer polymers induce the completely opposite state, immunosuppression. There is also some evidence of HA size being able to act as a switch between adhesive and migratory functions [44, 152]. If that is the case, the longer polymers and cable structures of HA are likely to induce the clustering of CD44s on the cell surface, and thereby alter the collective status of the CD44-HA interactions to favor signaling pathways inducing adhesion. Therefore, it is probable that at least some of the regulatory potential lies in the spatial arrangement and in the clustering of CD44, backed by the recent findings by Wolny et al. [149], who evaluated the effect of HA size and clustering on the CD44-HA interaction.

Wolny et al. [149] investigated two model systems: monomeric CD44 HABD and dimeric HABD. In both cases the CD44 HABD was fused with an immunoglobulin Fc domain and attached to an artificial membrane system. They found the binding of HA to increase in a sigmoidal fashion as a function of HA size, with a plateau after 262 kDa. The reversibility of the binding was also strongly dependent on HA size.

More precisely, when the molecular weight of the HA oligomers was below 10 kDa, the binding was reversible and highly dependent on HA concentration, whereas at molecular weights of over 30 kDa the binding was totally irreversible. This can be explained by the long HA chain having multiple "binding sites" for CD44s collectively rendering the adsorption irreversible. Interestingly, upon analyzing the packing of CD44 on the membrane surface they discovered the mean distance between individual HABDs to be twice as much as the peptide core of the protein. The authors therefore speculated that native glycosylation of the CD44 HABD might limit the packing of the proteins.

Undoubtedly, glycosylation plays a role in regulating the binding of an HA polymer to CD44 HABD. Supporting this claim, three different activity states for CD44 have been discovered from normal cell populations [9]. The first one is a constitutively active form constantly interacting with the HA polymers. The second form is inducible, so that unless activated by specific antibodies or cytokines, it does not bind to HA. The third form is constitutively inactive and cannot be activated even with the aforementioned factors. As the molecular weights of each of these forms were characterized with immunoprecipitation before and after deglycosylation, the inactive form was clearly the heaviest before the treatment, followed by the inducible form, and leaving the active form as the lightest. These differences were assigned specifically to different N-glycosylation patterns of CD44 HABD. In another study, inactive cells acquired the ability to bind HA after treatment with neuramidase, an enzyme cleaving off terminal sialic acids from the attached carbohydrates [153]. The concentration of CD44 remained unchanged through the experiment, implying that these qualitative changes might be more important for HA binding than the actual expression levels of the protein.

When regulating the CD44's binding affinity the most prominent target for glycosylation is, of course, HABD. It contains five N-glycosylation sites (Asn25, Asn57, Asn100, Asn110, and Asn120) from which two (Asn25 and Asn120) have been shown, through mutation experiments, to be especially sensitive to a glycosylation-induced disruption of the native CD44-HA interactions [15]. That is, deglycosylation of these residues has the ability to turn CD44 from an inducible to a constitutively active form in murine lymphoma cell lines [10]. The exact mechanism through which glycosylation mediates the CD44-HA interactions is unclear, but it can, in principle, occur in several ways including steric hindrance, charge repulsion, induction of conformational changes, or inhibiting the clustering of CD44. Interestingly, treatment with deoxyjirimycin, an antibiotic inhibiting the processing of N-linked glycans, did not transform the CD44 phenotype from inactive to active [154]. While opposing discoveries have also been made, the above finding could imply that alterations in HA binding stem from differential utilization of N-glycosylation sites rather than

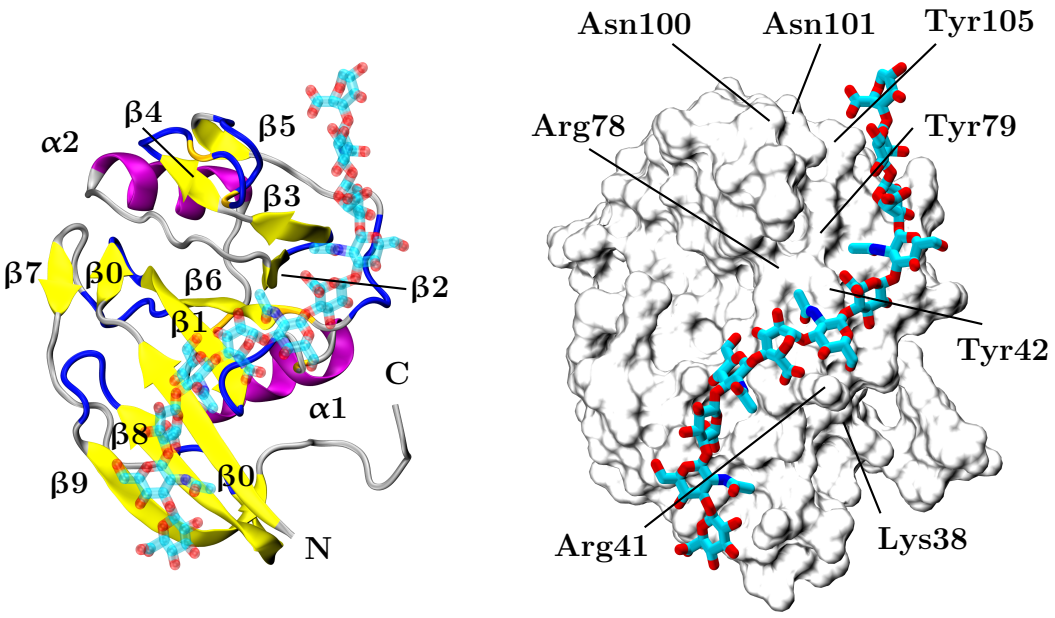


Figure 3.6: Structure of the CD44 HABD. The figure on the left shows a newcartoon representation of HABD colored according to the secondary structure. It also gives appropriate names for the secondary structure elements. The figure on the right shows a surface representation of HABD, where some of the relevant binding site residues are marked with arrows. The bound HA oligomer is colored cyan.

from the differential processing of the *N*-linked glycans. In another study, removal of some of the terminal sialic acids from the attached sugars was enough to transform CD44 from inactive to active [155], in turn implying that all the glycosylation sites are not equally important, as even the active form can still have some of these sites glycosylated.

3.4 Affinity for Carbohydrates Lies in the Hyaluronan Binding Domain

To understand the details of HA binding one must first understand the structure of CD44 HABD. The Link module is a highly conserved α/β-fold, with two antiparallel β-sheets consisting of six β-strands and two α-helices [15, 16]. In CD44, it is enlarged by so-called flanking N- and C-terminal sequences constituting four additional β-strands. The interacting residues lie mostly in the shallow groove located on the upper surface of CD44 as viewed in Figure 3.6. To shed light on the possible role of each *N*-glycosylation site, and to map the whole HA binding region, a range of structural studies have been performed over the last twenty years.

The earliest attempt to map the HA binding surface in CD44 involved sequencing of the primary structure of CD44 HABD [13]. The goal was to find areas of basic amino acids that would readily interact with the negatively charged hexuronate

groups of HA. As a result, two clusters were identified. The first one lies in the link homology region and involves amino acids 21 to 45. Importantly, supported by a mutation-induced loss of binding ability, arginine41 (Arg41) was found to be especially prominent in interacting with the HA oligomers. Later its role has been confirmed in multiple occasions [14, 16–19], with methods ranging from site-directed mutagenesis to computer simulations. The second cluster of basic amino acids, involving the most C-terminal amino acids of HABD (residues 144–167), was located outside the Link module. From these, Lys158 and Arg162 had the biggest contribution in binding the ligand, which has, however, later been questioned by Banerji et al. [16], who claim that the binding ability is conferred exclusively on the Link module.

After the first experiments by Peach et al. [13], Bajorath et al. [14] conducted a combined modeling and mutagenesis study to identify the precise residues participating in the HA binding. They built their model with homology modeling using a solution structure of TSG-6 as a template, and based on this, 24 residues were selected to be mutated. By comparing conservative and non-conservative mutations the authors were then able to deduce if the residues were directly or indirectly important for the binding of HA. For instance, amino acids Arg41, Tyr42, Arg78, and Tyr79, brought together by the β1-α1 and α2-β3 loops, form an especially important pincers adhering to the HA polymer and tying it firmly into place.

Teriete et al. [15] solved the first crystal structure of a non-glycosylated human CD44 HABD. A comprehensive analysis of the structure verified many of the earlier results obtained by Bajorath et al., including the role of most of the binding residues. Furthermore, additional NMR studies suggested that the Link module was stabilized upon HA binding, while the C-terminal flanking sequences of HABD were subjected to a major conformational change unfolding the last residues of HABD. This is also consistent with the NMR findings by Takeda et al. [19, 156, 157], who interpreted the attained chemical shift patterns of C-terminal residues 152 to 169 to be a clear indication of a rather large conformation transition from ordered (O) to a partially disordered (PD) state. Encouraged by these findings, the same group conducted an experimental *in vitro* assay where they observed a dynamic balance between O- and PD-conformations, with PD being the higher affinity state [158]. Intriguingly, they also proposed that the coexistence of these two conformations could in macroscopic scale enable the leukocytes to roll in a velcro-like fashion. However, this idea has not been verified by any experimental means.

Meanwhile, Banerji et al. [16] presented a crystal structure of mouse CD44 in complex with an HA octamer in two different conformations. Interestingly though, these two conformations were totally unrelated to the ones found by Teriete et al. and Takeda et al. Instead of a huge relaxation of the C-terminus, these changes

were related to the position of the Arg41 side-chain. The ligand was bound in both conformations, but the so-called A-form had no apparent contact between HA and the side-chain of Arg41. By contrast, in B-form the Arg41 containing loop was tilted slightly to allow the side-chain to flip towards the bound HA and to establish hydrogen bonds between each other. As the previous crystal structure without the bound HA was in A-form, the authors suggested that the B-form corresponds to a high affinity state.

Two MD studies then addressed the effect of these conformational transitions [17, 18]. First, Jamison et al. [17] identified the typical Arg41 side-chain rotamer populations and torsion angles of the Arg41-containing loop characteristic to the A- and B-forms. As a conclusion the authors claimed that the otherwise unfavorable B-form becomes highly favorable in the presence of an HA polymer. They also measured the free energies of these two conformations to be roughly equal. A very recent study by Favreau et al. [18], however, found the free energy of the B-form to be less than that of the A-form's. The authors therefore assigned the B-form to represent the most favorable state in the presence of a bound ligand. They also measured the free energy of A- and B-forms in both O- and PD-conformations and, surprisingly, found no evidence of allostery between these two conformational changes. However, they did propose structural evidence for the high affinity binding in the PD state. That is, as conformational flexibility of the C-terminal tail in the PD conformation is increased, Arg154 is, consistent with early finding by Peach et al., able to form more favorable interactions with the bound HA oligomer.

4. COMPUTATIONAL METHODS AND SIMULATION PARAMETERS

Nowadays, computer simulations play a highly important role in science. Yet, in the past, science was based on two corner stones: theory and experiment [22]. Theory often expresses the problem in a form of a mathematical equation. This kind of approach is, however, reasonable only for a limited number of simplified cases where the mathematical formulation yields a strict analytical solution. Today, theories are more often based on the dynamics of a vast number of particles, and deriving exact analytical solutions to such systems is impossible without a considerable amount of simplification.

Similarly, with experiments one has to handle large sets of research data gathered from different measurements. This data is then fitted to an existing theory or sometimes novel features are discovered through generalization and imagination. Still, the complexity of the results is often overwhelming and analytical solutions that fit the measurement data are difficult to find [159]. Furthermore, experiments are still rather limited in their resolution. In other words, nanoscale phenomena, like the interactions between biomolecules, are too rapid or tiny to be measured accurately with experiments.

As computers have become more common and sufficiently powerful, they have become a prominent tool in scientific research [22]. In fact, today computer experiments or simulations are being regarded as the third corner stone of science. The models for computer simulations are still provided by experiments and theory, but the calculations are executed numerically by a machine that is following a certain simulation algorithm. As a result, the investigated systems can be more complex and realistic. Indeed, simulations allow one to understand the experimental results in detail even at the nanometer scale. In addition, simulations can be employed to carry out experiments too hazardous or expensive to perform in a real world [159].

It is important to remember that the results from computer simulations are always approximations and simplifications of nature. For this reason, these results are eventually always confirmed with experimental studies with real life physical models. Strangely enough, oversimplification is in fact a true advantage of computational methods. By discarding some of the irrelevant information, systems under study can be reduced to a level where one can actually understand the physical laws

governing a given process. That is basically the fundamental idea of modeling.

This chapter introduces the basic concepts of molecular dynamics simulations. First, it gives an overview of the topic. Then, in the first and second subsections, it describes the principle of accounting for the chemical and physical bonds between atoms, and discusses the fundamental idea behind performing these simulations. The next subsection then addresses the fundamental approximations employed in MD. The second section reviews the theory of obtaining free energies from MD simulations. Lastly, this chapter is closed by describing the simulated systems and listing all the simulation parameters.

4.1 Molecular Dynamics for Simulating Biomolecules

Molecular dynamics (MD) is a computational method employing classical Newtonian mechanics to model the time evolution of a given group of objects [160]. MD enables the simulation of discrete particles as small as individual atoms. Thus, it is a perfect method for studying biological molecules and their dynamic behavior with atomistic precision. MD is able to predict equilibrium properties of a classical many-body system. The term 'classical' means that electronic degrees of freedom are neglected in MD, and thus the motions of the nuclei obey the laws of classical mechanics [161]. In other words, some accuracy is discarded at the expense of reaching large enough time scales. Examples of phenomena studied with MD include the permeation of drug molecules through lipid membranes [162], the collective motions of lipids in a membrane [163], and the complex interplay between lipids and protein molecules [164].

The simulated systems under study always comprise just a fraction of the whole cell or a cell membrane. This is because larger ones would be impossible to study with the current computer resources. However, the model systems still consist of up to several million atoms, and simulating them requires a great deal of time and computational resources. Therefore, large supercomputers and computer clusters are often employed in order to reduce the required wall-clock time and to reach the desired level of statistical accuracy. On the other hand, by simplifying the systems it should be possible to achieve longer simulation time scales, but then some microlevel phenomena might go undetected. Indeed, simulating biological systems through model studies requires one to balance the available resources and the desired accuracy of the results. At the same time it is worthwhile to realize that an artificial physical model may produce misleading results even in the absence of any computational limitations.

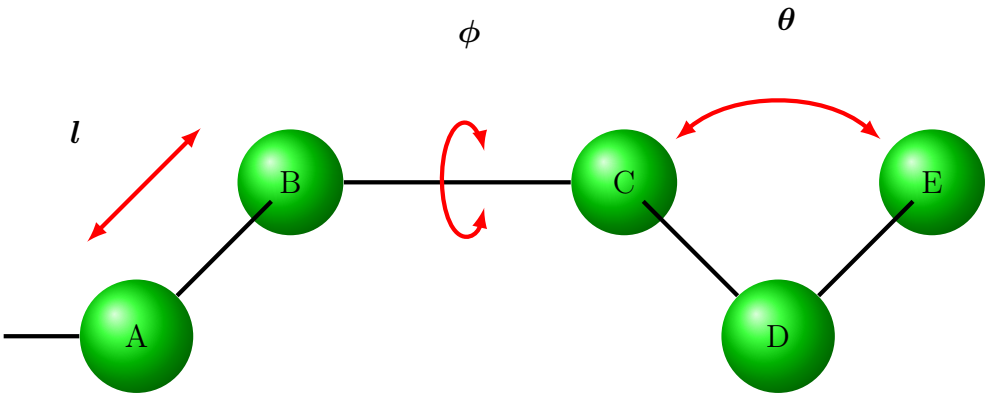


Figure 4.1: A schematic illustration of a molecule and its internal degrees of freedom in MD. The bond stretching motion l is depicted as the relative movement between atoms A and B. Dihedral angle rotation ϕ is the angle between planes formed by atoms ABC and BCD, while θ marks the bending of the angle formed by atoms CDE.

4.1.1 Force Field Describes Molecular Interactions

To understand the principle of MD one has to first consider intra- and intermolecular interactions [161]. In classical MD the electronic distribution is approximated into the form of an analytical potential energy function also known as the Hamiltonian operator $\mathcal{H}_{\text{total}}$. It describes all the molecular interactions classically, by dividing them into two categories, bonded interactions and non-bonded interactions:

$$\mathcal{H}_{\text{total}} = \mathcal{H}_{\text{bonded}} + \mathcal{H}_{\text{non-bonded}}. \tag{4.1}$$

Bonded interactions of a molecule refer to the internal degrees of freedom or simply to the relative motion of the constituent chemical bonds. They comprise bond stretching, angle bending, and dihedral rotation which are all further clarified in Figure 4.1. Bonded interactions are taken into account with a generic potential energy term of the following form:

$$\begin{aligned}\mathcal{H}_{\text{bonded}} = &\sum_{\text{bonds}} \frac{1}{2} k_{l,ij}(l_{ij} - l_{ij,\text{eq}})^2 + \sum_{\text{angles}} \frac{1}{2} k_{\theta,ijk}(\theta_{ijk} - \theta_{ijk,\text{eq}})^2 + \\ &\sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi_{ijkl} - \phi_{ijkl,\text{eq}})],\end{aligned} \tag{4.2}$$

where the terms describe bond stretching, angle bending, and dihedral angle rotation, respectively. Parameters $k_{l,ij}$ and $k_{\theta,ijk}$ are the force constants for bond stretching and angle bending, while l_{ij} and $l_{ij,\text{eq}}$ represent the bond length between atoms i and j and its corresponding equilibrium value. Similarly, values θ_{ijk} and $\theta_{ijk,\text{eq}}$ describe the angle between atoms i, j, and k, and its equilibrium value. Dihedrals are expressed with a Fourier cosine series where n is the periodicity, V_n is

the corresponding Fourier component, and ϕ_{ijkl} is the dihedral angle along with its equilibrium position $\phi_{ijkl,\,\mathrm{eq}}$. Values for k's are usually markedly high, so that bonds will not overstretch unnaturally but rather stay at realistic conformations.

In MD, all non-bonded interactions are described with pairwise additive potentials. In practice, the non-bonded Hamiltonian contains three terms

$$\mathcal{H}_{\mathrm{non-bonded}} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right) + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \right) f_{ij}, \tag{4.3}$$

which describe short-range steric repulsion, van der Waals interactions, and long-range Coulombic interactions, respectively. The first two interactions are combined into the Lennard-Jones potential, where the ϵ_{ij} parameter is related to the energy scale of the interaction describing the depth of the potential well. Meanwhile, σ_{ij} is related to the length scale of the interaction and is therefore called the van der Waals distance. Parameter r_{ij} denotes the distance between interacting atoms i and j. Strangely enough, the exponent of the repulsive term has no physical justification, but is chosen only for the sake of computational efficiency as it is calculated just by squaring the attractive term. Yet, it performs desirably well in describing the interaction.

While the Lennard-Jones potential applies to all the particles in the system, Coulombic interaction is calculated only between those particles i and j that bear a charge q_i and q_j, respectively. The magnitude of these charged interactions depend on the surrounding media according to ε_0 and ε_r that describe the permittivity of vacuum and the relative permittivity of the surrounding solvent, respectively. Importantly, both Lennard-Jones and Coulombic interactions are scaled with a scaling factor f_{ij}, so that $f_{ij} = 0$ for atoms that are less than three bonds apart in the same molecule. In comparison, the scaling factor gets a value of 0.5 when the atoms are exactly three bonds apart from each other, and a value of 1.0 in all the other cases.

In addition to the functional form, another important aspect in describing the molecular interactions is the parameterization of the potentials. However, there is no systematic way to parameterize these functions. Because of the relatively simple form of the interaction potentials, each particle type requires different arguments. Usually parameterization is conducted by comparing simulation results with *ab initio* quantum-mechanical calculations or with experiments. Unfortunately, the various parameters between different atoms are often interdependent, and therefore finding a functional and reliable combination requires continuous iteration.

Together, the potential function and its parameterization are called the force field. Despite its name the forces acting on each particle can be extracted only after differentiating the potential function. Today, there are several force fields for the

use of MD with a slightly different parameterization in each, and none of them is able to describe all the physical properties of every atom type. However, they are continuously being refined and thus, several versions of the same force field often exist [161].

In this Thesis, we employed a force field called AMBER [165]. Its name stands for Assisted Model Building with Energy Refinement, and we used the refined version called AMBER ff99SB-ILDN [166]. It is a commonly used force field that is especially good for describing proteins and nucleic acids, but it has also several different parameter sets for other biomolecules. One that is particularly important for this Thesis is the GLYCAM parameter set [167] for simulating carbohydrates such as GAGs.

4.1.2 Simulation Trajectory Evolves Through Numerical Integration

The MD algorithm itself is rather straightforward. A simplified depiction of it is presented in Figure 4.2. Importantly, this sequence of events is essentially the same regardless of what simulation platform is being applied. For example, in this Thesis we employ the GROMACS [161, 168, 169] simulation package, which is an open-source software package supporting several different force fields.

To start an MD simulation one needs to construct topologies for all the constituent particles. They define the constant force field parameters and the structure of each atom type. Next, one must choose a correct statistical ensemble, such as constant volume or constant pressure, which usually reflects the appropriate conditions in related experiments. Before initiating the simulation, one still needs to define the timescale it will cover and how frequently it will be updated. Today, typical timescales reach up to several microseconds while the interval of one simulation time step is usually only a few femtoseconds. The initial structure of the system is finally built by inserting the desired particles along with solvent molecules and ions into a simulation box. This box can adopt multiple shapes, but the square is by far the most practical due to its simplicity.

After having defined the system, all the constituent particles are assigned initial velocities [161]. These are in turn generated from the Maxwell-Boltzmann distribution that reads

$$p(v_i) = \sqrt{\frac{m_i}{2\pi k_\mathrm{B} T}} \exp\left(- \frac{m_i v_i^2}{2 k_\mathrm{B} T} \right), \tag{4.4}$$

where k_B is the Boltzmann constant, v_i and m_i are the speed and mass of the particle i, and T is the absolute temperature. Importantly, the starting configuration is never in equilibrium. It must, however, be reasonably close to it, so the system

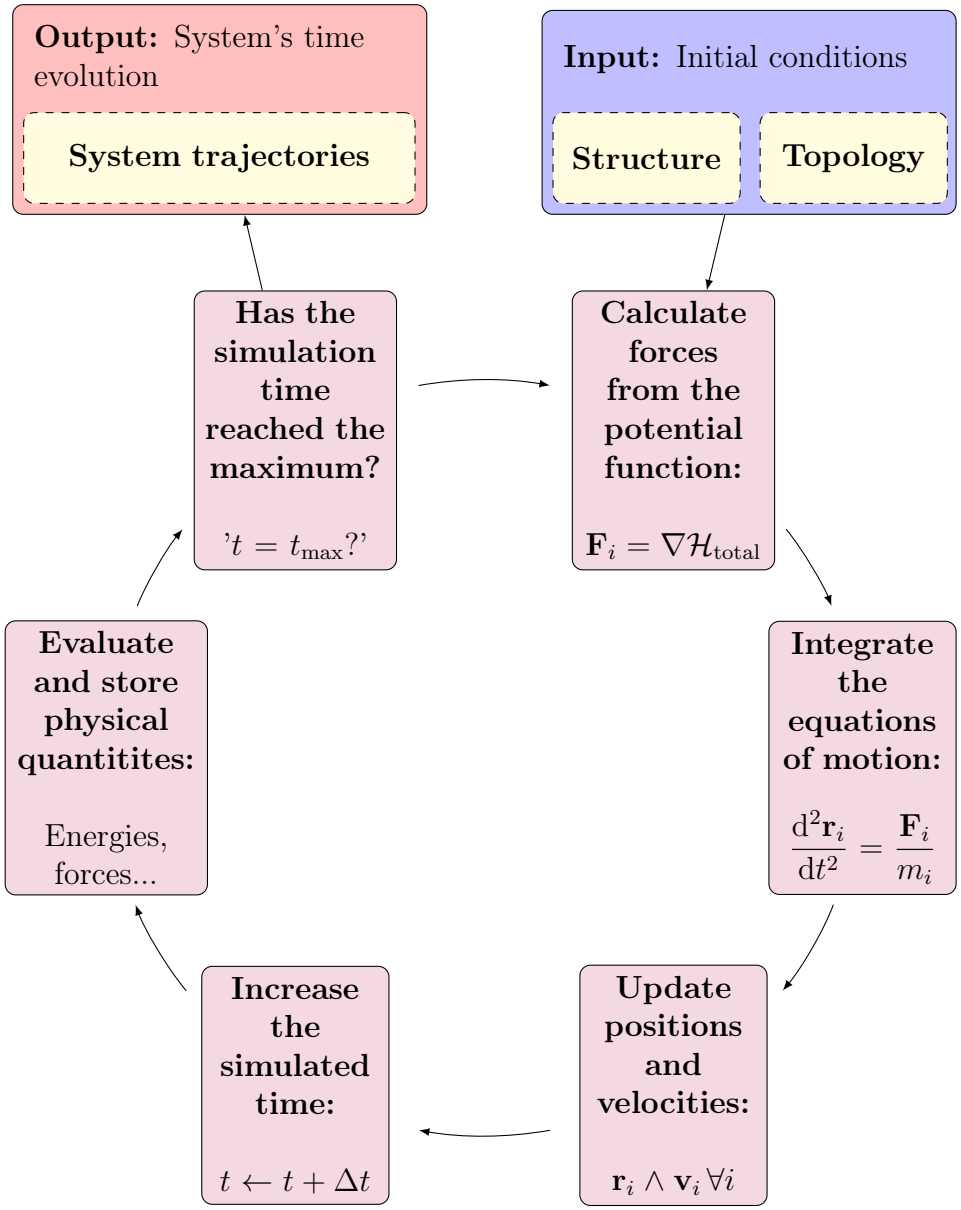


Figure 4.2: Schematic illustration of the MD algorithm. $\mathbf{F}_i$ expresses the forces acting on particle i, $\mathcal{H}_{\text{total}}$ is the total potential function, $\mathbf{r}_i$ and $\mathbf{v}_i$ give the position and velocity of particle i, respectively, and similarly m_i describes the mass of particle i. Parameter t states the current simulation time, Δt is the timestep, and $t_{\max}$ tells the endpoint of the simulation.

is able to find a stable state after a short energy minimization. In other words, a specific minimization algorithm, like steepest descent or conjugate gradient, probes the system's potential energy surface by moving the constituent atoms slightly with respect to their initial positions in zero temperature. This procedure ensures that none of the particles is subjected to unnaturally high forces when the actual simulation commences.

The basis of MD lies in the numerical integration of the equations of motion.

Through this procedure the algorithm updates the position and velocity of each particle at every timestep. However, to be able to solve the equations of motion, the forces acting on each atom are first required. They are obtained at every simulation step before the actual integration by taking the negative gradient of the potential function with respect to distances and angles to other particles (Figure 4.2). Consequently, the force calculation is by far the most computationally demanding part of the MD algorithm.

Once the forces are finally derived, there are several implementations that can perform the numerical integration. A particularly efficient one is called the leap-frog algorithm [170]. As the name vaguely suggests, the algorithm calculates positions and velocities in turns, so that at every half step the velocity is solved as

$$\mathbf{v}_i\left(t+\frac{\Delta t}{2}\right)=\mathbf{v}_i\left(t-\frac{\Delta t}{2}\right)+\frac{\mathbf{F}_i(t)}{2m_i}\Delta t, \tag{4.5}$$

while the positions are solved at whole steps

$$\mathbf{r}_i(t+\Delta t)=\mathbf{r}_i(t)+\mathbf{v}_i\left(t+\frac{\Delta t}{2}\right)\Delta t. \tag{4.6}$$

Parameters $\mathbf{v}_i$ and $\mathbf{r}_i$ in Eqs. (4.5) and (4.6) denote the velocity and the position of particle i, respectively. Parameter $\mathbf{F}_i(t)$ expresses the force acting on particle i at time t, while t and Δt are the actual simulation time and timestep, respectively.

By repeating this general sequence countless times the simulation algorithm gradually produces the output trajectories describing the system's evolution as a function of time. Eventually, the desired properties, like the energetics related to ligand binding, can be extracted from these trajectories with a myriad of different analyzing techniques.

4.1.3 Computational Shortcuts Save Time and Resources

As any computational simulation, also MD employs several shortcut schemes. These approximations are intended to improve the simulation's runtime behavior by reducing the number of required force calculations, and in some cases to render the system in many ways more realistic. For example, MD employs cut-offs to keep the potential short-ranged, and thus to retain the force calculations at minimum. This means that by truncating or shifting the potential, it will reach the value zero after a certain cut-off threshold distance r_{cut}. Such a procedure is particularly suitable for van der Waals interactions since they decay fast as a function of distance. All in all, the cut-off scheme causes a small bias in the potential energy of the system, but in turn reduces the computational complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$, where N is the

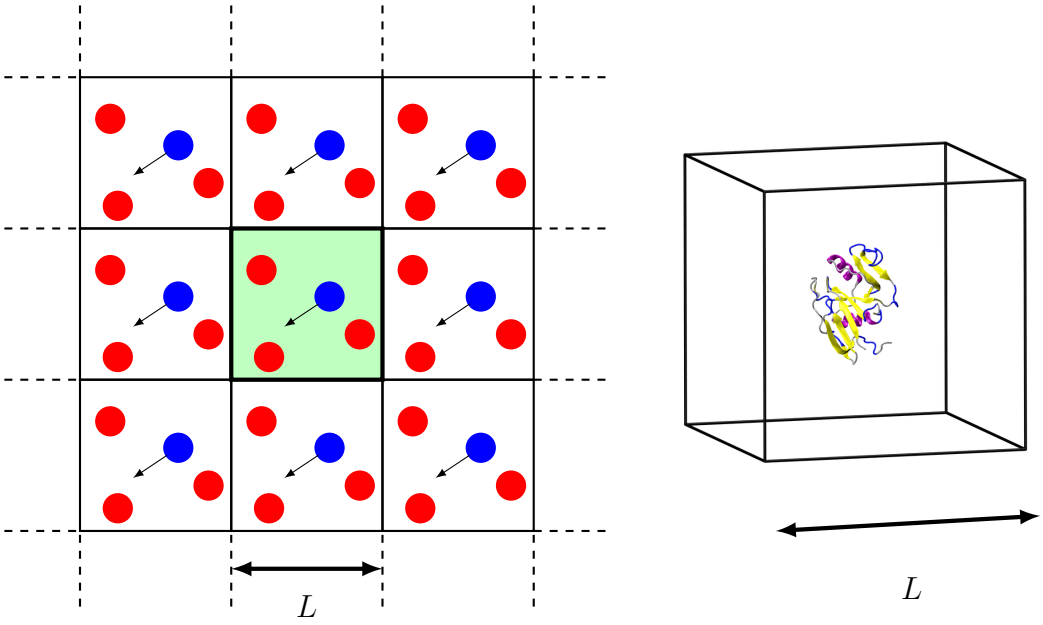


Figure 4.3: On the left, a schematic figure representing the fundamental idea of periodic boundary conditions in 2D. The unit cell along with the particle trajectories are copied in all directions and if a particle tries to leave the cell it simply emerges back from the opposite side. The parameter L is the size of an individual cell. On the right, a real unit cell is given as a snapshot from real simulation, where the CD44 protein is placed at the center of the simulation box. Water molecules are omitted from this figure for clarity.

number of particles in the system [31]. Here we stress that r_{cut} is used for van der Waals interactions, but typically not for other more long-ranged interactions (see below).

As a consequence of the cut-off, it will suffice to calculate the forces only between neighboring particles. Therefore, the simulation box is geometrically divided into smaller cubic cells whose size is only slightly larger than the cut-off radius. Then it will be sufficient to evaluate the forces between the particles in the neighboring cells only. As a result, this neighbor list is updated at defined intervals.

Furthermore, to avoid unwanted surface effects, a scheme called periodic boundary conditions is employed. This means that the simulation box is periodically repeated in every direction, as illustrated in Figure 4.3. If a particle leaves the unit cell through the left border, it therefore immediately emerges back from the right border. Importantly, each particle i interacts with the closest periodic image of the other $N-1$ particles even if they were in the adjacent cell. As a result, one is able to simulate the bulk properties of the system with just a small number of particles. However, great care must be taken to choose the cut-off distance properly, so that particles do not interact with their own periodic images. The r_{cut} is consequently at most $L/2$, where L is the size of the cubic simulation box [31].

One problem arising from the cut-offs used for van der Waals interactions is the

mapping of the long-range electrostatic interactions. Unlike the Van der Waals potential, electrostatic interactions decay slowly and thus cannot be neglected at distances larger than r_{cut}. Fortunately, there is a number of ways able to tackle this problem. Today, the most accurate and computationally efficient technique is the so-called Particle Mesh Ewald (PME) method [171, 172].

When including the periodic images, the electrostatic potential of N interacting particles is given by the following Ewald sum:

$$\mathcal{H}_{\mathrm{Ewald,Coulomb}} = \frac{1}{2}\sum_{i}^{N}\sum_{j}^{N}{\sum_{|\mathbf{n}|}}' \frac{q_i q_j}{|\mathbf{r}_{ij}+\mathbf{n}|}, \tag{4.7}$$

where $\mathbf{r}_{ij}$ is the interparticle distance vector from atom i to j and $\mathbf{n}$ is the box index vector $\mathbf{n} = (n_x L, n_y L, n_z L)$, where the n_x, n_y, and n_Z are integers. Accordingly, the denominator represents the real distance between charges i and j. The prime symbol in the sum indicates that $\mathbf{n} = 0$ when $i = j$, so that the denominator is well defined. Ewald's trick [173] was to render this slowly converging sum into two rapidly-converging terms combined with a correction term

$$\mathcal{H}_{\mathrm{Ewald,Coulomb}} = \mathcal{H}_{\mathrm{Ewald,real}} + \mathcal{H}_{\mathrm{Ewald,recip}} + \mathcal{H}_{\mathrm{Ewald,corr}}, \tag{4.8}$$

where the first term is evaluated in real space while the second is solved in reciprocal space. Following certain algebraic manipulations, the real space sum is given by

$$\mathcal{H}_{\mathrm{Ewald,real}} = \frac{1}{2}\sum_{i,j}^{N} q_i q_j {\sum_{|\mathbf{n}|}}' \frac{\mathrm{erfc}(\beta|\mathbf{r}_{\mathrm{ij}}+\mathbf{n}|)}{|\mathbf{r}_{ij}+\mathbf{n}|}, \tag{4.9}$$

where the complementary error function erfc is defined as

$$\mathrm{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)\mathrm{d}t. \tag{4.10}$$

Meanwhile, the reciprocal sum can be expressed as

$$\mathcal{H}_{\mathrm{Ewald,recip}} = \frac{1}{2\pi L^3}\sum_{i,j}^{N} q_i q_j \sum_{|\mathbf{m}|\neq 0} \frac{\exp(-\pi^2|\mathbf{m}|/\beta^2)}{|\mathbf{m}|^2}\exp[2\pi \mathrm{i}\mathbf{m}\cdot(\mathbf{r}_{\mathrm{i}}-\mathbf{r}_{\mathrm{j}})], \tag{4.11}$$

and the final correction term then reads

$$\mathcal{H}_{\mathrm{Ewald,corr}} = \frac{-\beta}{\sqrt{\pi}}\sum_{j=1}^{N} q_j^2, \tag{4.12}$$

where $\mathbf{m}$ is a reciprocal box vector and β determines the relative weight of the

real and reciprocal sums. In real space, the sum is calculated directly according to the cut-off while the reciprocal sum is spread onto a discrete grid with multidimensional piecewise interpolation. Fortunately, reciprocal energy and force terms are expressed as convolutions and can therefore be evaluated quickly using the Fast Fourier Transform (FFT) algorithm. As a result of applying FFT, the PME method provides a significant improvement to the simulation algorithm's runtime behavior when calculating electrostatic interactions [31].

Similarly, controlling the proper statistical ensemble requires the exploitation of certain computational shortcuts. By default, the MD algorithm produces a microcanonical NVE (constant number, constant volume, and constant energy) ensemble. Yet, biological systems often allow volume and energy to fluctuate while the temperature and pressure are instead kept constant. Reaching these isothermal-isobaric (NpT) conditions through simulations therefore requires the utilization of thermo- and barostats forcing the temperature and pressure to stay constant (within thermodynamic fluctuations) during the run.

Like with many schemes discussed above, also thermo- and barostats have multiple implementations to choose from. For instance, in this Thesis we applied the ones called the velocity rescaling (v-rescale) thermostat [174] and the Parrinello–Rahman barostat [175]. The former controls the temperature by rescaling the velocities of the constituent particles with a properly chosen stochastic factor. This ensures a correct kinetic energy distribution

$$\mathrm{d}E_{\mathrm{kin}} = (E_{\mathrm{kin},0} - E_{\mathrm{kin}})\frac{\mathrm{d}t}{\tau_t} + \sqrt{\frac{E_{\mathrm{kin}}E_{\mathrm{kin},0}}{N_{\mathrm{f}}}}\frac{\mathrm{d}W_w}{\sqrt{\tau_t}}, \tag{4.13}$$

where $E_{\mathrm{kin},0}$ is kinetic energy of the coupled thermal reservoir, E_{kin} is the current kinetic energy, τ_t is the time constant, N_{f} is the number of degrees of freedom, and $\mathrm{d}W_w$ is a Wiener process. The advantage of the v-rescale thermostat is the first order decay of temperature deviations and the lack of oscillations, while still producing a correct statistical ensemble. Similarly, the Parrinello–Rahman barostat gives the true NpT ensemble. It controls the pressure $\mathbf{P}$ by allowing the volume V of the box to fluctuate according to the matrix equation of motion

$$\frac{\mathrm{d}\mathbf{b}^2}{\mathrm{d}t^2} = V\mathbf{W}^{-1}\mathbf{b}'^{-1}(\mathbf{P} - \mathbf{P}_{\mathrm{ref}}), \tag{4.14}$$

where $\mathbf{b}$ is the matrix representing the box vectors, $\mathbf{P}_{\mathrm{ref}}$ is the reference pressure while $\mathbf{W}$ is the matrix parameter that determines the strength of the coupling and how strongly the box can be deformed. Its inverse is defined as

$$\mathbf{W}^{-1} = \frac{4\pi^2\alpha_{ij}}{3\tau_p^2 L_{\max}}, \tag{4.15}$$

where α_{ij} is the isothermal compressibility, τ_p is the time period of pressure fluctuations, and $L_{\max}$ is the largest box matrix element. If applied when system is not properly equilibrated, the Parrinello–Rahman coupling may result in overly high box fluctuations, yet it is a reliable choice for an MD barostat.

In MD, certain constraints are usually applied to augment the equations of motion by freezing the highest-frequency interactions. With fastest vibrations removed from the model, the simulation timestep can be increased considerably, thus allowing one to explore longer timescales. Furthermore, fixing the bond lengths does not significantly alter the dynamical properties of the system. Fixing the intraparticle bond angles, on the other hand, often leads to biased dynamics. Nowadays, the most common algorithm to constrain non-solvent molecules is called LINCS (Linear Constraint Solver) [176], whereas another, faster but less flexible one, named SETTLE [177], is applied to constrain solvent molecules like water.

4.2 Umbrella Sampling Method for Measuring Free Energies

The calculation of free energy differences with MD simulations is a key application in computational biophysics [178–180]. This is because free energy differences are the driving forces for every physical or chemical process, such as the adsorption of a ligand to its receptor. Consequently, partitioning and mean position of such a ligand can be described with a free energy profile which, in theory, could be obtained from unbiased MD simulations [180]. In practice, however, the non-ergodicity of those systems severely hinders the sampling of the process at reachable timescales. This means that states with low probability (i.e. high energy) are not sampled sufficiently, as the probabilities for such processes are obtained according to the Boltzmann factor

$$\exp(-A/k_{\mathrm{B}}T), \tag{4.16}$$

where A is the Helmholtz free energy. For example, if the barrier A is much higher than 10 $k_{\mathrm{B}}T$, the statistical probability for a spontaneous process at a physiological temperature becomes extremely low.

Umbrella sampling [181] is a method where system's ergodicity is expanded artificially by the use of harmonic restraining potentials [178]. These potentials act along a certain reaction coordinate ξ to drive the system from a thermodynamic state A to B. The reaction path is then divided into N_w intermediate steps, as shown in Figure 4.4 which illustrates the basic concept of umbrella sampling. In practice, each of these steps is an individual sampling window where an MD simulation is conducted in order to obtain the biased distribution of the system along the reaction coordinate ξ_i ($i = (1, \ldots, N_w)$). These windows are forced to overlap, so that the individual distributions can later be unbiased and combined to a single global

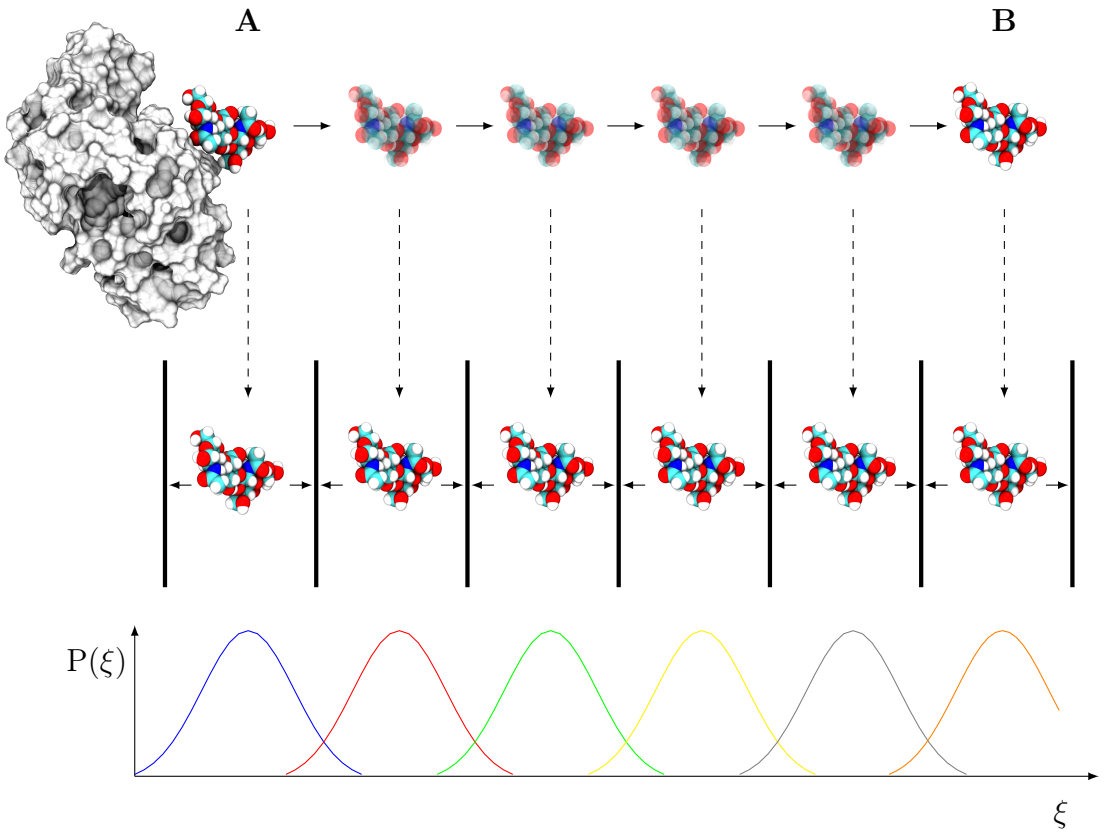


Figure 4.4: The principle of umbrella sampling. As the system passes from a thermodynamic state A to state B along a reaction coordinate ξ, such as in this figure where the ligand molecule (HA) is pulled away from its binding site on the surface of a protein (CD44 HABD), its displacement trajectory is sliced into separate sampling windows. When simulated individually with restraints, these sampling windows produce a biased sampling distribution each representing a small part of the process. The free energy is then calculated based on these overlapping distributions.

distribution based on which the final free energy profile is then calculated from.

Theoretically, obtaining the free energy profiles from MD simulations is not a straightforward task. In describing the state of a system with N particles, a total of $6N$ dimensions are required to describe the so-called phase space of the system [182]. That is, every particle is described with three spatial and three momentum coordinates. However, free energy is not simply a function of these phase space coordinates of the system, but rather related to the canonical partition function Q [178].

The Q can be calculated by integrating over the whole phase space [178]. If the potential energy E is independent of momentum, Q is obtained simply as

$$Q = \int \exp[-\frac{1}{k_{\mathrm{B}}T}E(r)]\mathrm{d}^{N_{\mathrm{f}}}r. \tag{4.17}$$

In a canonical ensemble, the number of particles, volume, and temperature are kept

fixed, and as a result the Helmholtz free energy is related to Q via

$$A = -k_{\mathrm{B}}T\,\mathrm{ln}Q. \tag{4.18}$$

If pressure, on the other hand, is kept constant instead of temperature, the Gibbs free energy is obtained. This is relevant especially in biological systems, as water is non-compressible. Therefore, in MD simulations of biological systems, ΔA and ΔG are numerically almost indistinguishable.

In many cases the thermodynamic states A and B are separated by a simple reaction coordinate ξ. In principle, it can be defined in several ways, such as a change in the Hamiltonian. Most often it is, however, a geometric parameter like distance or torsion between two atoms or groups of atoms [178]. In this Thesis, it was chosen to be the distance between the centers-of-mass of CD44 and the HA oligomer along one coordinate in the Cartesian space.

When ξ is defined, all the other degrees of freedom can be integrated out of Eq. (4.17) to obtain a probability distribution of the system along ξ:

$$Q(\xi) = \frac{\int \delta[\xi(r) - \xi]\mathrm{exp}[-\frac{1}{k_{\mathrm{B}}T}E]\mathrm{d}^{N_{\mathrm{f}}}r}{\int \mathrm{exp}[-\frac{1}{k_{\mathrm{B}}T}E]\mathrm{d}^{N_{\mathrm{f}}}r}, \tag{4.19}$$

where $Q(\xi)\mathrm{d}\xi$ is the probability of the system to reside in a short interval $\mathrm{d}\xi$ around ξ. However, the problem is that in computer simulations direct phase space integrals like Eq. (4.19) are impossible to calculate. Fortunately, if the system is assumed to be ergodic, like in umbrella sampling, the ensemble average $Q(\xi)$ is equal to the time average $P(\xi)$. As a result, $A(\xi)$ also known as the potential of mean force (PMF) is according to Eq. (4.18) defined as

$$A(\xi) = -k_{\mathrm{B}}T\,\mathrm{ln}P(\xi), \tag{4.20}$$

and it is thereby directly obtainable from computer simulations [178].

To ensure the sampling of all regions along the reaction coordinate ξ, a biasing potential w_i is applied in every simulation window i to hold the system in a given position [178]. Importantly, w_i is an additional energy term and depends only on ξ:

$$E^{\mathrm{b}}(r) = E^{\mathrm{u}}(r) + w_i(\xi), \tag{4.21}$$

where 'u' and 'b' refer to unbiased and biased quantities, respectively. The most

common $w_i(\xi)$ is a harmonic biasing potential of the following form:

$$w_i(\xi) = \frac{K_i}{2(\xi - \xi_i^{\mathrm{ref}})^2}, \tag{4.22}$$

which restraints the system at a position ξ_i^{ref} $(i = 1, \ldots, N_w)$ with a force constant K_i $(i = 1, \ldots, N_w)$, thereby producing the biased distribution of the system along the reaction coordinate ξ in each window N_w. The choice of K is of crucial importance in umbrella sampling, as too low K will cause the system to avoid the low probability regions whereas too high K will not produce sufficient overlap between the individual distributions.

The unbiased free energy $A_i(\xi)$ is obtained from the unbiased probability distribution in each window. According to Eq. (4.19), the latter is written as

$$P_i^{\mathrm{u}}(\xi) = \frac{\int \delta[\xi'(r) - \xi]\exp[-\frac{1}{k_{\mathrm{B}}T}E]\mathrm{d}^{N_{\mathrm{f}}}r}{\int \exp[-\frac{1}{k_{\mathrm{B}}T}E]\mathrm{d}^{N_{\mathrm{f}}}r}. \tag{4.23}$$

Biased distribution is, on the other hand, in accordance with Eq. (4.21) written as

$$P_i^{\mathrm{b}}(\xi) = \frac{\int \delta[\xi'(r) - \xi]\exp[-\frac{1}{k_{\mathrm{B}}T}(E(r) + w_i(\xi'(r)))]\mathrm{d}^{N_{\mathrm{f}}}r}{\int \exp[-\frac{1}{k_{\mathrm{B}}T}(E(r) + w_i(\xi'(r)))]\mathrm{d}^{N_{\mathrm{f}}}r}. \tag{4.24}$$

Because the bias caused by $w_i(\xi)$ depends only on ξ and the integration is performed over all degrees of freedom except ξ, Eq. (4.23) and Eq. (4.24) can be combined into one relation:

$$P_i^{\mathrm{u}}(\xi) = P_i^{\mathrm{b}}(\xi)\exp[\frac{1}{k_{\mathrm{B}}T}w_i(\xi)]\langle\exp[-\frac{1}{k_{\mathrm{B}}T}w_i(\xi)]\rangle. \tag{4.25}$$

Now, by realizing that $P_i^{\mathrm{b}}(\xi)$ is obtained from MD simulations and that $w_i(\xi)$ is given analytically, the unbiased free energy for the ith window is

$$A_i(\xi) = -k_{\mathrm{B}}T\ln P_i^{\mathrm{b}}(\xi) - w_i(\xi) + F_i, \tag{4.26}$$

where the undetermined constant F_i represents the free energy related to applying the window potential [178]. It is defined from the result known as the Jarzynski equality [183]:

$$\exp[-\frac{1}{k_{\mathrm{B}}T}F_i] = \langle\exp[-\frac{1}{k_{\mathrm{B}}T}w_i(\xi)]\rangle, \tag{4.27}$$

which, given enough sampling, is sufficient to unbias the simulation as long as it

spans the whole ξ [178].

In addition to just unbiasing, the individual distributions must also be combined in order to be able to compute the global PMF. This means that the constant F_i needs to be calculated and, as it is not directly obtainable from sampling, special methods are required. From these, the most established one is called the Weighted Histogram Analysis Method (WHAM) [184]. It offers a way to estimate F_i and, at the same time, to minimize the statistical error of $P^{\mathrm{u}}(\xi)$ through an iterative process. The central idea is that the global distribution is based on the weighted average of the individual window distributions:

$$P^{\mathrm{u}}(\xi) = \sum_{i}^{N_w} p_i(\xi) P_i^{\mathrm{u}}(\xi), \tag{4.28}$$

where p_i denotes weights assigned to the individual distributions. They are chosen to minimize the statistical error of P^{u} in such a way that

$$\frac{\partial \sigma^2(P^{\mathrm{u}})}{\partial p_i} = 0, \tag{4.29}$$

where normalization is done such that $\sum p_i = 1$. As a result,

$$p_i = \frac{a_i}{\sum_j a_j},\ a_i(\xi) = N_i \mathrm{exp}[-\frac{1}{k_{\mathrm{B}}T} w_i(\xi) + \frac{1}{k_{\mathrm{B}}T} F_i], \tag{4.30}$$

where N_i denotes the total number of simulation steps in the window i. The constant F_i is, in turn, calculated by integrating the right hand side term from Eq. (4.27) over ξ to obtain

$$\begin{aligned} \mathrm{exp}[-\frac{1}{k_{\mathrm{B}}T} F_i] &= \int P^{\mathrm{u}}(\xi) \mathrm{exp}[-\frac{1}{k_{\mathrm{B}}T} w_i(\xi)] \mathrm{d}\xi \\ &= \int \mathrm{exp}[-\frac{1}{k_{\mathrm{B}}T} (A(\xi) + w_i(\xi))] \mathrm{d}\xi. \end{aligned} \tag{4.31}$$

Now, by iterating Eq. (4.28) and Eq. (4.31) for P^{u} and F_i until convergence, Eq. (4.26) can be applied to construct the global PMF profile. The speed of the convergence, however, depends on the number of umbrella windows and on the height of the existing free energy barriers [178].

On a more practical side, Hub et al. [185] introduced a WHAM implementation called g_wham that works with the GROMACS simulation software package. In addition to evaluating the PMF profiles of the system in an above-mentioned way, it offers estimates of statistical errors by employing a technique called the bootstrap analysis [186].

The principle of the bootstrap analysis is to generate hypothetical observations

using the existing umbrella histograms $h_i(\xi)$ as a template [185]. In other words, bootstrapped trajectories $\xi_{b,i}(t)$ for each umbrella histogram $h_i(\xi)$ are obtained based on existing probability distributions of ξ_i $(i = 1, \ldots, N_w)$. These virtual trajectories then form a new set of N_w histograms $h_{b,i}$ that are subsequently given as input to WHAM yielding a bootstrapped PMF $\mathcal{W}_b(\xi)$. After repeating this procedure N_b times, a total of N_b bootstrapped PMFs $\mathcal{W}_{b,k}(\xi)$ $(k = 1, \ldots, N_b)$ is acquired. Notably, in this study, N_b was set to a value of 200.

As a result of having multiple trajectories, the error in the PMF is evaluated simply by taking the standard deviation of the bootstrapped PMFs with

$$\sigma_{PMF}(\xi) = [(N_b - 1)^{-1} \sum_{k=1}^{N_b} (\mathcal{W}_{b,k}(\xi) - \langle \mathcal{W}_b(\xi) \rangle)^2]^{1/2}, \tag{4.32}$$

where $\langle \mathcal{W}_b(\xi) \rangle = N_b^{-1} \sum_{i=k}^{N_b} (\mathcal{W}_{b,k}(\xi)$ is the average of the bootstrapped PMFs at the position ξ. If the sampling in each window N_w is limited due to relatively long autocorrelation times, such as in conformational sampling of macromolecules, it is recommended to perform multiple simulations of each window with a random seed. However, when sampling is sufficient, the bootstrapping procedure is accurate and gives reasonable error estimates [185].

4.3 Model Systems and Simulation Parameters

In this study, the simulated systems consisted of differentially glycosylated CD44 proteins and HA oligosaccharides with various sizes. The system was initially constructed by downloading the crystal structure of a human CD44 HABD by Teriete et al. [15] from the Protein Data Base with the identifier 1UUH. After this, the following steps were taken. First, a second monomer and all the water molecules were removed from the model in order to have a system of one CD44 protein. Second, the 1UUH structure contained two selenomethionines, artificial amino acids that help the structure elucidation of proteins in x-ray crystallography, at positions 63 and 66 that needed to be replaced with normal methionine residues. The proper topology was then created with the GROMACS tool pdb2gmx, after which the protein was placed at the center of a simulation box of 8 nm × 8 nm × 8 nm in size.

The topologies of the HA oligomers were generated based on GLYCAM06 [167] force field parameters. We abbreviate the HA oligomers used in this study as HA_4, HA_6, HA_8, and HA_{16}, with the suffix denoting the number of monosaccharides in a given fragment. In order to build these HA oligomers, a special program was created. It reads a residue database of monosaccharides and subsequently concatenates the carbohydrate monomers in a certain conformation. The program also adds the correct end groups (hydroxyl groups and hydrogen atoms) to the termi-

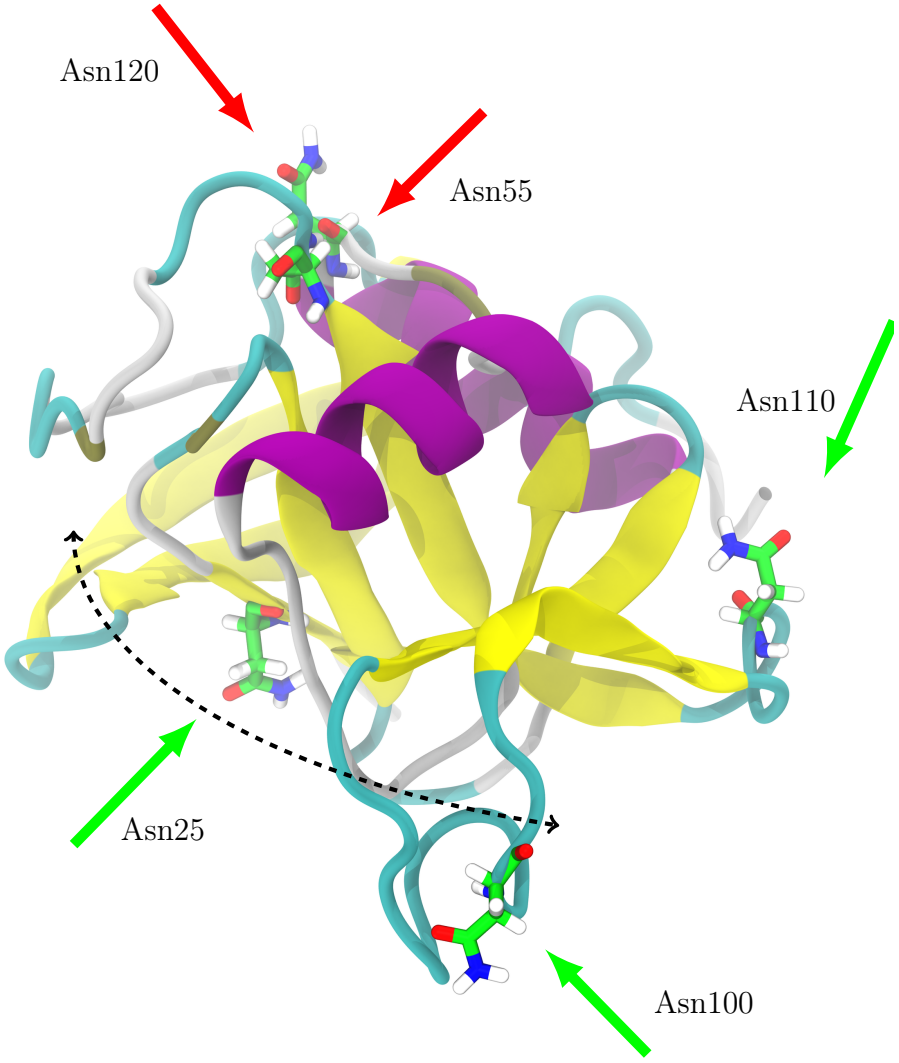


Figure 4.5: Amino acid residues selected for glycosylation. The selection was made based on the proximity to the binding site. All in all, Asn25, Asn100, and Asn110 were glycosylated, each in separate systems, while Asn57 and Asn120 were left unmodified. The dashed arrow denotes the HA binding surface.

nal monosaccharides. Before incorporation into the final model, the ready-made HA fragments were still geometry optimized with a program called Avogadro [187], employing steepest the descent algorithm and the MMFF94 force field.

CD44 was glycosylated by covalently adding core pentasaccharides to its surface. Three of the five potential *N*-glycosylation sites (Asn25, Asn100, Asn110) were chosen to be glycosylated, so that only one carbohydrate was added per simulation system. The selection of the *N*-linkage sites was made based on the proximity to the HA binding groove, as illustrated in Figure 4.5. Residues Asn57 and Asn120 were left unglycosylated because they reside on the opposite side of the protein and are therefore less likely to directly affect the binding of HA oligomers. The pentasaccharides were also built with a custom-made program. It attaches the sugar

residues to a given Asn residue of the protein and provides the user with a correct topology file. Notably, one has to be aware that AMBER and GLYCAM06 force fields do not share the same scaling factors. Because proteins are very sensitive to changes in the scaling factors, the modifications had to be subjected to the sugar parameters. The automatic topology builder program for glycosylated proteins was therefore programmed to convert the pentasaccharide topology parameters to match AMBER scaling factors. Importantly, this and all the aforementioned programs are available at Ref. [188].

Given the large variance in the size, glycosylation scheme, and conformational status of the studied HA and CD44 structures, multiple systems were simulated. Table 4.1 lists all of these simulations and designates which modifications were made or observed in their structures. In all of the systems the protein was initially placed at the center of the simulation box, after which water and salt (NaCl) were added to give the physiological concentration (150 mmol/l) of salt in each system, thereby reflecting the conditions inside the extracellular fluid, such as blood plasma. In addition to this, extra Na^+ ions were added to render the system electrically neutral, as both CD44 and HA bear a negative total charge. Lastly, before conducting any actual production simulations the systems were carefully energy minimized.

All the simulations were performed with the GROMACS version 4.6.3 [161]. The initial energy minimization was always conducted with the following sequence of steps. First, each system went through 1000 steps of steepest descent minimization with a 1 fs timestep. To ensure that the system was properly balanced, this was followed by 1000 steps of conjugate gradient minimization with a 2 fs timestep. In production simulations the integration was performed by the leap-frog algorithm with a 2 fs integration timestep. Furthermore, in contrast to the minimization runs where there were no bond constraints applied to non-water molecules, the production simulations utilized the LINCS algorithm to keep the bond distances fixed. The TIP3P water model and the SETTLE algorithm were employed for modeling and constraining the water molecules, respectively.

The simulations were performed under periodic boundary conditions in all directions along with PME electrostatics to take care of the long-range electrostatics. A cutoff distance of 1 nm was used with the short-ranged Lennard-Jones interactions and neighbor lists. Furthermore, all simulations were conducted in the NpT ensemble, so that the v-rescale thermostat was employed to couple the systems to a heat bath that was set to a physiological human body temperature, 310 K, with a time constant of 0,1 ps. Separate heat baths were used for the biomolecules under study and the salt water surrounding them. Lastly, the Parrinello–Rahman barostat coupled the systems isotropically to a reference pressure of 1 bar, with a time constant of 1 ps and a compressibility of $4.5 \times 10^{-1}\,\mathrm{bar}^{-1}$.

Table 4.1: Simulations in this Thesis. The first column gives an abbreviation for each simulation. The second column describes the systems' contents. Importantly, if a spontaneous flip to the B-form conformation was observed, it is designated in brackets. Furthermore, free energy calculations are denoted as "Umbrella", and if the protein was glycosylated it is marked with the name of the attachment residue. Lastly, the third column gives the duration of the simulation trajectory, and the fourth column reports the size of the simulation box.

Name	Description	Duration (ns)	Size (nm)
S_1	CD44 with HA_{16} 1	1000	$9 \times 9 \times 9$
S_2	CD44 with HA_{16} 2 (B-form)	1700	$10 \times 10 \times 10$
S_3	CD44 with HA_{16} 3	1000	$10 \times 10 \times 10$
S_4	CD44 with HA_{16} 4	1000	$10 \times 10 \times 10$
S_5	CD44 with HA_8	1000	$9 \times 9 \times 9$
S_6	CD44 glycosylated at Asn25	500	$8 \times 8 \times 8$
S_7	CD44 glycosylated at Asn100 (B-form)	1000	$8 \times 8 \times 8$
S_8	CD44 glycosylated at Asn110	500	$8 \times 8 \times 8$
S_9	CD44 glycosylated at Asn25 +HA_{16}	1000	$11 \times 11 \times 11$
S_{10}	CD44 glycosylated at Asn100 +HA_{16}	1000	$11 \times 11 \times 11$
S_{11}	CD44 glycosylated at Asn110 +HA_{16}	1000	$11 \times 11 \times 11$
S_{12}	Umbrella: CD44 +HA_4	19×100	$9 \times 9 \times 9$
S_{13}	Umbrella: CD44 +HA_8 1	34×100	$10 \times 10 \times 10$
S_{14}	Umbrella: CD44 Asn25 +HA_8 outwards	36×100	$10 \times 10 \times 10$
S_{15}	Umbrella: CD44 Asn25 +HA_8 inwards	25×100	$10 \times 10 \times 10$
S_{16}	Umbrella: CD44 +HA_8 2	32×100	$10 \times 10 \times 10$
S_{17}	PD-CD44	500	$10 \times 10 \times 10$
S_{18}	PD-CD44 +HA_8 2	500	$10 \times 10 \times 10$

5. SIMULATION RESULTS REVEAL HOW CD44 IS REGULATED

This chapter presents the results of this study, compares them to existing literature, and discusses their biological significance. For extracting relevant information from the simulation trajectories we used both ready-made GROMACS analysis tools [161] and our own data-extraction programs. The latter were mainly scripted with the Python 2.7 programming language [189], but also with the MATLAB® numerical computing environment. All the snapshot pictures were visualized with VMD [35] and rendered with the Tachyon ray tracing library [190].

The main objective of this Thesis is to explore the effects of HA size, CD44 glycosylation, and CD44 conformation transitions on the protein-carbohydrate interaction. Accordingly, this chapter is divided into three sections each describing one of these topics. The first section compares the residual binding profiles for HA_8 and HA_{16} and presents the binding free energy profiles for HA_4 and HA_8.

The second section reports the results for the glycosylated proteins. First, the viability of our models is verified by comparing the N-linkage dihedral angles to known literature. Again, residual binding profiles show how each glycosylation changes the adsorption pattern of HA_{16}. Free energy profiles then present compelling thermodynamic evidence of how a glycosylation inside the binding groove affects the binding process.

The third section discusses the two previously characterized conformation changes of the CD44 HABD. First we assess the enhancement of binding related to the unfolding of the C-terminal tail and compare our findings to a previous computational study [18]. We also report our somewhat provocative findings regarding the sharp and local A-to-B transition. Importantly, the structure of our protein model is initially in the O-state (C-terminus is confined) and in the A-form (Arg41 side-chain is open). Therefore, all the following simulation results are obtained, unless reported otherwise, for this conformation.

5.1 Ligand Size Affects the Strength of Binding

To measure the contribution of each CD44 amino acid to the binding of the HA oligomer, we calculated the average hydrogen bonds and minimum distances between HA_{16} and every residue of the protein from simulations S_1, S_2, S_3, and S_4. For this

purpose, we used two geometric criteria to define the hydrogen bonds. The distance cut-off between the hydrogen and the donor atom was set to 0.35 nm, and the angle cut-off in the triplet formed by the hydrogen, acceptor, and donor atoms was limited to 30°. In Figure 5.1, the calculated profiles are compared with similar graphs derived from simulation S_5, where HA_{16} was replaced by a two times shorter HA_8.

Collectively, the data shows three regions important for the binding. First, the core of the binding groove is formed by Arg41, Lys78, and Arg79, with Arg41 making two constitutive hydrogen bonds with the bound HA. Second, residues 108 to 113, located in the loop between the β5 and β6 strands, appear to have a relatively high contribution in the binding, forming on average five hydrogen bonds with the bound HA oligomer. Third, the Arg154 and the other residues from the C-terminal end seem to have an effect only when the ligand is long enough. The last observation is particularly interesting because at least two of the previous studies claim [16, 18] that the contribution from Arg154 is negligible, while demonstrating this only with HA_8. By comparison, our results indicate that at least HA_{16} is capable of forming an average of one hydrogen bond with Arg154 and thus implying that the interaction is energetically meaningful.

After clearly defining the principal binding site, we carried out two umbrella sampling simulation sets to measure the effect of HA chain length on the strength of the interaction at this particular region. It was done as follows. First, the starting structures for the steered MD simulations were taken from snapshots of spontaneously bound CD44-HA complexes. After this, a harmonic biasing potential was applied to pull the center of mass (COM) of a stably bound HA away from the binding groove. Being more specific, 40 ns simulations were used to pull the HAs (HA_4 and HA_8) along the reaction coordinate with a force constant of $1000\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ and a pulling rate of $0.1\,\mathrm{nm\,ns^{-1}}$. Furthermore, the reference pull group was selected to be the backbone of residues 75–79. These amino acids are buried inside the "back wall" of the binding groove and their fluctuations are therefore minimal compared to other residues at the same site. As a result, they impose very little statistical error to the subsequent free energy calculations.

To ensure that the pulling of HA was executed only in one direction, and to significantly reduce the complexity of the process, the protein was restrained in space, so that the binding groove was directly facing the x-axis (the reaction coordinate ξ), as illustrated in Figure 5.2. The restraining was, in essence, conducted by choosing three benchmark points from the protein backbone (C_α atoms from residues Thr47, Phe56, and Gln65) to be subjected to a restraining potential. This choice was, in turn, made based on two criteria. First, the atoms had to be located as far from the binding groove as possible to avoid skewing the interactions with the bound HA.

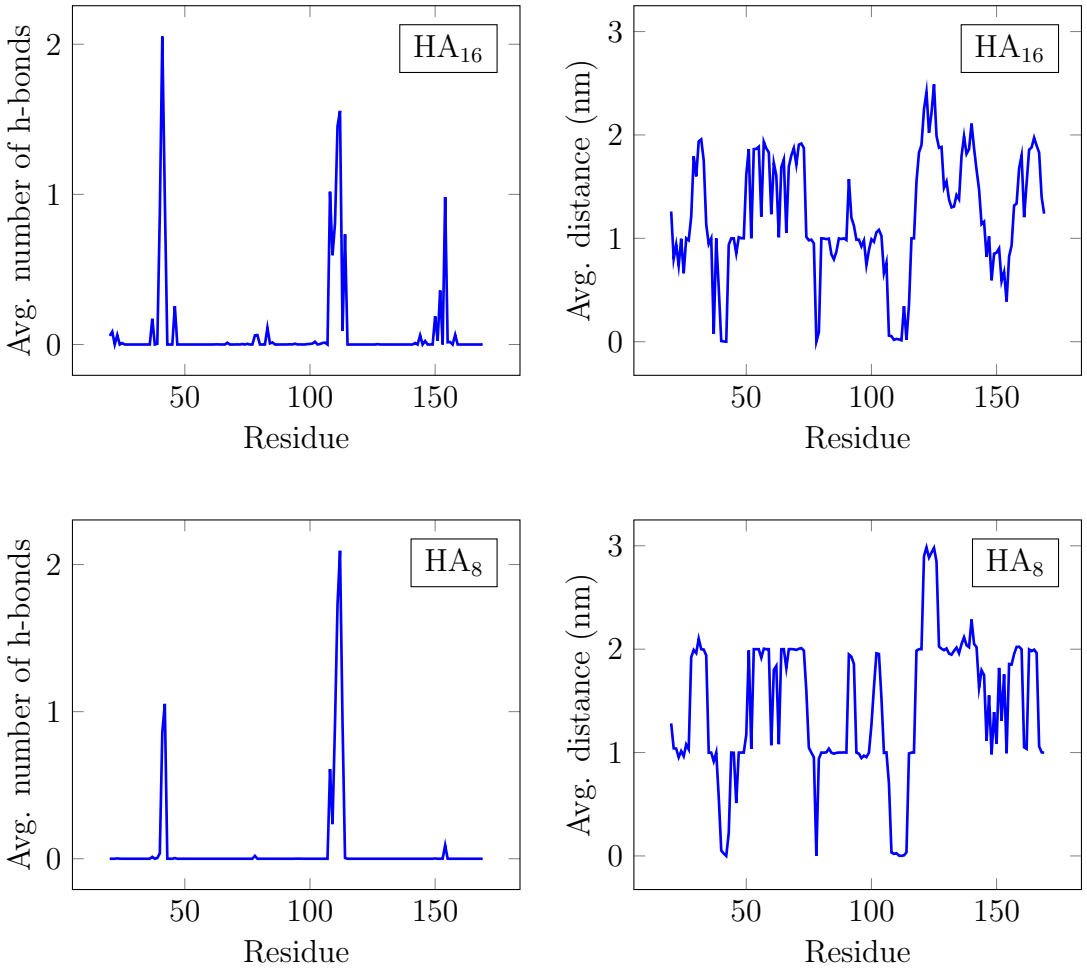


Figure 5.1: Hydrogen bond and minimum distance profiles per residue for two cases, with HA_8 and HA_{16}, as marked in the legends of the figures. The values for HA_{16} are averaged from simulations S_1, S_3, and S_4, where HA was stably bound and the Arg41 side-chain was in the A-conformation through the simulations. The profiles for HA_8 are calculated from the simulation S_5. The first 50 ns were discarded from the analysis in each case.

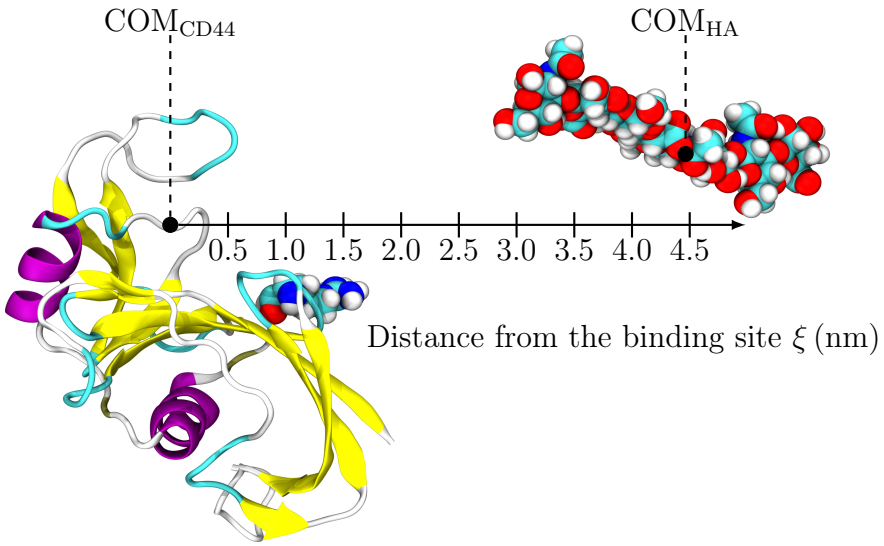


Figure 5.2: A snapshot figure representing the simulation setup. The protein is restrained into this position by freezing the positions of three backbone atoms (atom numbers 347, 559, and 692). Furthermore, the reaction coordinate is the distance from the back of the binding site (the backbone atoms of residues 75 to 79). The Arg41 residue in the border of the binding site is shown in order to clarify the orientation of the protein.

Second, the atoms had to have relatively low root-mean-square fluctuations, so that restraining them would impose only a minimal bias to the overall structure of the protein.

Starting structures for umbrella sampling windows were then collected as snapshots from the pulling simulation and spaced 0.10 ± 0.01 nm apart along the x-axis. These were chosen to achieve equal sampling along the whole reaction coordinate ξ. In each window, the HA oligomer was constrained only in x-direction and was thus allowed to move freely in the corresponding yz-plane. The pull groups and the force constants were exactly the same as in the previous pulling simulations. From the 100 ns sampling windows, only 80 ns was used for the actual free energy calculation, as the first 20 ns were regarded as equilibration, and therefore cropped from the real analysis. The final free energy profiles and their statistical error estimates are presented in Figure 5.3.

The results confirm our previous observations. In both cases the free energy minimum lies at a distance of 0.9 to 1.3 nm from the reference pull group, which corresponds to the position of the Arg41 (Figure 5.2). From the two ligands, the HA tetramer has a binding free energy ΔG_b of roughly -7.5 kJ mol^{-1} ($\approx 3\,k_\mathrm{B}T$ in a temperature of 310 K). This binding energy is in the order of thermal fluctuations and the ligand is therefore able to make only transient interactions with CD44. Now, it is worthwhile to realize that to be able to collectively facilitate, for instance, the

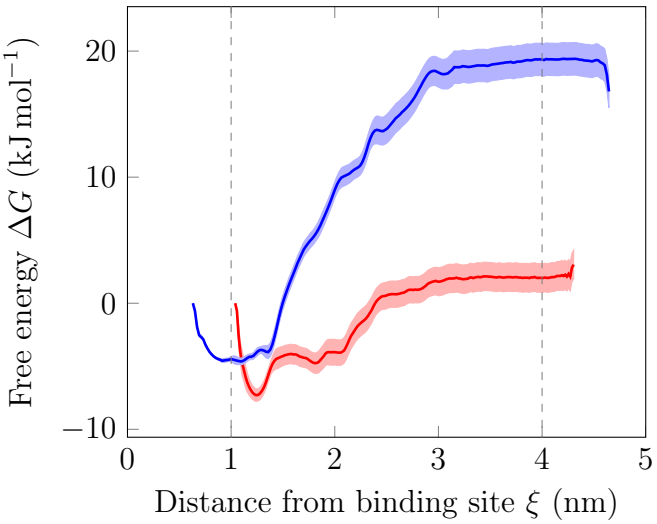


Figure 5.3: Free energies for the binding of HA tetramer and octamer. The free energies from simulation S_{12} are given in —, and from simulation S_{13} in — with their error limits shaded accordingly. The horizontal axis is equivalent to the one in Figure 5.2 and the horizontal dashed lines at 1 nm and 4 nm mark the positions of the Arg41 side-chain and the threshold for bulk water, respectively.

rolling of the leukocytes, the individual CD44-HA interaction needs to be reversible (i.e. close to the energy scale of thermal fluctuations). Our results, however, show that the binding between the HA octamer and CD44 is, at least in the absence of any external forces such as the drag caused by the blood flow, rather irreversible with ΔG_b of -25 kJ mol^{-1}. This finding therefore suggests that some dynamic mechanism is required to control the binding affinity of longer HA polymers.

The binding free energy has also recently been estimated with wet-lab experiments. Banerji et al. [191] measured the dissociation constant k_D of individual CD44-HA interaction, using a surface plasmon resonance and non-linear curve fitting to Langmuir isotherm, to be 65.7 µM. Fortunately, the dissociation constant is related to ΔG via

$$\Delta G = RT \ln\left(\frac{k_\mathrm{D}}{c^\ominus}\right), \tag{5.1}$$

in which R is the ideal gas constant and $c^\ominus = 1.0\ \mathrm{mol\,L^{-1}}$ is the standard reference concentration. By plugging in the values we can estimate that

$$\begin{aligned}\Delta G &= 8.314\,\mathrm{J\,mol^{-1}\,K^{-1}} \times 310\,\mathrm{K} \ln\left(\frac{65.7 \times 10^{-6}\,\mathrm{mol\,L^{-1}}}{1\,\mathrm{mol\,L^{-1}}}\right) \\ &= -24820.84512\,\mathrm{J\,mol^{-1}} \\ &\approx -24.8\,\mathrm{kJ\,mol^{-1}},\end{aligned}$$

which is remarkably close to the approximate height of the free energy barrier for the adsorption of HA_8 that we calculated with umbrella sampling. It should, however, be

Table 5.1: Angles in the carbohydrate-protein attachment. Asn25, Asn100, and Asn110 designate the amino acid residue of the *N*-linkage.

Angle or dihedral	Asn25	Asn100	Asn110	Reference value [33]
O5-C1-Nδ1	103.68	104.862	105.356	106.8
C2-C1-Nδ1	113.319	112.955	113.101	111.1
C1-Nδ1-Cγ	123.618	123.813	122.802	121.4
O5-C1-Nδ1-Cγ	-113.528	-90.2792	-90.7735	-100.0
C1-Nδ1-Cγ-Cβ	115.988	26.3762	-74.3934	-179.0
Nδ1-Cγ-Cβ-Cα	-139.379	19.6102	-103.687	-172.0
Cγ-Cβ-Cα-N	96.9236	-80.5216	-72.1692	-69.0

noted that their result is measured for a high molecular weight HA and is therefore not strictly comparable to our findings.

5.2 Glycosylations Modulate Native Binding Interactions

We modeled glycosylations to study how the covalently attached carbohydrates on the surface of the protein regulate the binding affinity of the HA oligomers. At first, we carried out three 500 ns simulations (S_6, S_7, and S_8), with the glycosylated proteins, but without any ligand, to see how the pentasaccharides would orient themselves relative to the body of the protein and what effect does their presence have on its stability. Fortunately, previous analyzes of *N*-linkages on glycoproteins have provided information about the stereochemistry at the carbohydrate-protein attachment. So, the average angles and dihedrals from the *N*-linkages are tabulated and compared to this reference data in Table 5.1. As the data shows, the bond angles are remarkably well in line with the literature values. In comparison, the four dihedrals which determine the orientation of the attaching GlcNAc with respect to the local protein backbone exhibit more variance when compared to the reference values. Analysis of the previous x-ray data has, however, revealed that it is common for these dihedral angles to have broader distributions that vary according to the surroundings of the attachment site [33]. Furthermore, our measurements of root-mean-square deviations (RMSD) indicate that the structural stability of CD44 is not affected by the pentasaccharides (data not shown).

The final frames from simulations S_6, S_7, and S_8 are presented as snapshots in Figure 5.4. These pictures show that, in their final configurations, Asn100 and Asn110-linked pentasaccharides protrude into the water phase, whereas the Asn25-linked one prefers to clutch to the side of the protein forming a lid over the binding groove. It thereby gives a plausible explanation for why the glycosylation of Asn25 has been observed to modulate the binding of HA in a particularly discrete manner [10].

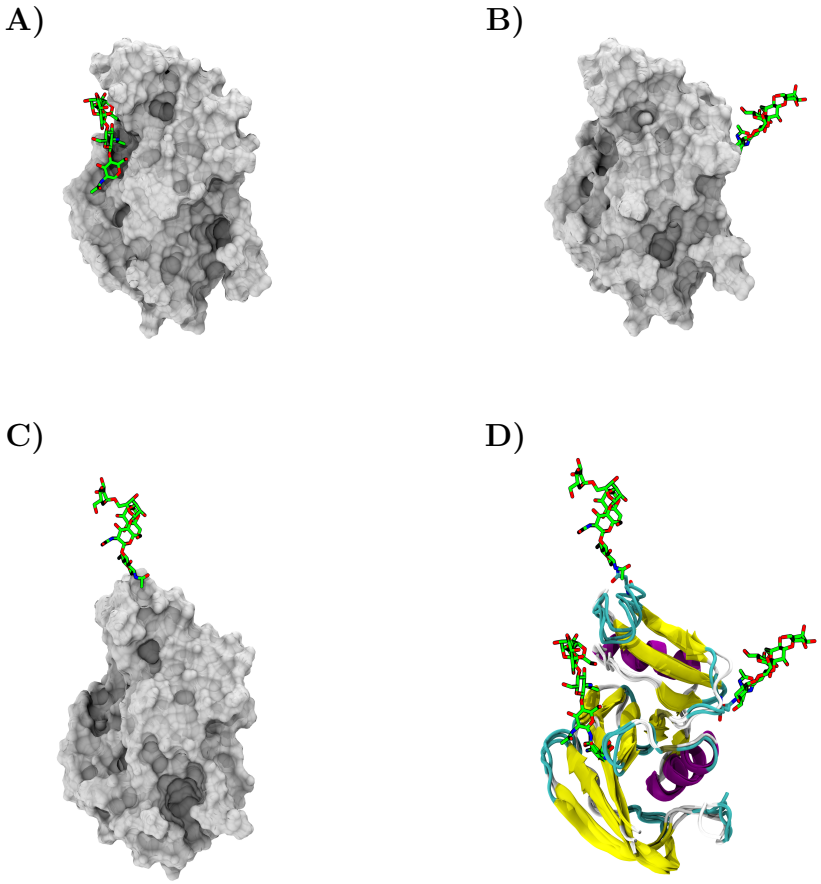


Figure 5.4: CD44 HABD with the three glycosylation schemes. A), B), and C) show the ending frames from the simulations S_6, S_7, and S_8 as snapshot pictures, where the protein is illustrated with a surface and the attached carbohydrates with a licorice presentation. Panel D shows the three structures aligned on top of each other. The surrounding salt water and hydrogen atoms of the carbohydrates are omitted for clarity.

To further test this hypothesis, we measured the thermodynamic effect of the presence of a pentasaccharide in Asn25 to the binding of HA octamer. For this purpose, the method of choice was again the umbrella sampling technique. Similarly, the orientation of the protein and the selection of the simulation parameters were mostly the same as shown in Figure 5.2 and as described in the previous section, respectively. The main exception was, however, that now the bound HA was placed in between the protein and the pentasaccharide (simulation S_{14}). We recognize that this configuration is probably not a naturally-occurring one, but nevertheless it can act as a thermodynamic reference configuration. Now it was also possible to use exactly identical atomic coordinates for both systems, with and without the added pentasaccharide (simulation S_{16}) and thus gain more comparable data. Overall, the free energy calculation proceeded with the following sequence of steps. First, the HA_8 was pulled outwards from the binding groove so that it was forced to bend the attached carbohydrate on its way. After this, we took a frame from the middle of the pulling simulation where HA had already released itself and was now facing the pentasaccharide from the side of the water phase. The frame then served as a starting structure for an opposite pulling simulation, this time forcing HA back to the binding groove that was now occupied solely by the pentasaccharide (simulation S_{15}).

For measuring the free energy, the pulling simulations were divided into a series of umbrella sampling windows. To achieve sufficient sampling of the whole reaction coordinate, even when pushing towards the pentasaccharide, the harmonic force constant K was set to 2000 $\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ in sampling windows 1–6 of the simulation S_{14} and in every window of the simulation S_{15}. Similarly, in order to have more accurate estimates of the position of the ligand, the pull group was changed to be just the ring atoms of the fourth carbohydrate unit in the HA octamer.

The free energy profiles along with their statistical error estimates are presented in Figure 5.5. In simulation S_{14} (red graph) the free energy difference from the binding groove to the bulk water is roughly -55 $\mathrm{kJ\,mol^{-1}}$ with a sharp minimum at 0.8 nm from the reference point, designating that when HA is artificially trapped to this position the native binding interactions are enhanced. Most importantly, the core pentasaccharide at Asn25 causes steric hindrance preventing the trapped HA from detaching. Similarly, in simulation S_{16}, in the system without the obstructing pentasaccharide, the HA octamer finds a strong free energy minimum from the binding groove (blue graph in Figure 5.5). This time the binding free energy is about -33 $\mathrm{kJ\,mol^{-1}}$, which is slightly higher than in a corresponding simulation S_{13}. The reason for this discrepancy lies most likely in the different starting frames, where the position of the bound ligand was not identical, in turn implying that distinct ligand binding modes may exist even within the same binding site.

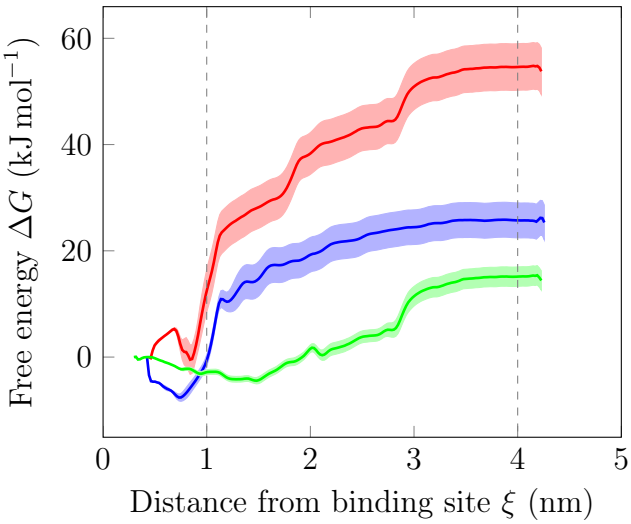


Figure 5.5: Free energies for the binding of the HA octamer to the glycosylated CD44. The free energies from the simulation S_{14} are given in —, simulation S_{15} in —, and simulation S_{16} in — with their error limits shaded accordingly. The horizontal axis is again equivalent to the one in Figure 5.2 and the horizontal dashed lines at 1 nm and 4 nm mark the positions of the Arg41 side-chain and the threshold for bulk water, respectively.

While in the simulation S_{14} the binding grove is clearly the global minimum along the reaction coordinate, the simulation S_{15} shows that in a more realistic situation where the HA octamer is attempting to dock into the glycosylated binding groove, the attached pentasaccharide effectively blocks most of the native binding interactions. In other words, the inward-out and outward-in processes for ligand binding are fundamentally different when the protein is glycosylated at Asn25. In a normal adhesion event, once HA approaches the binding groove, the positively-charged Arg41 starts to attract it intensively, as illustrated by a very steep free energy gradient in the distance interval of 0.8–1.1 nm. However, the absence of this gradient in the free energy curve corresponding simulation S_{15} (green graph in Figure 5.5) means that the steric repulsion of the core pentasaccharide prevents the ligand from aligning correctly into the binding groove. As a result, a shallow free energy minimum lies now at 1.5 nm, outside the reach of the Arg41 side-chain. All in all, the strength of the binding, roughly -19.5 kJ mol^{-1}, is reduced by 40 % from the unglycosylated case.

The above conclusion is further backed by the hydrogen bond and minimum distance profiles shown in Figure 5.6. They are calculated from the simulations S_9, S_{10}, and S_{11}, so that the last residues, ranging from 170 to 174, represent the five carbohydrate monomers of the attached pentasaccharide. Upon analyzing these graphs, it is apparent that glycosylating Asn25 will moderately decrease the overall number of hydrogen bonds between the ligand and the protein, at least when compared to the non-glycosylated hydrogen bond profiles in the previous section, whereas the other

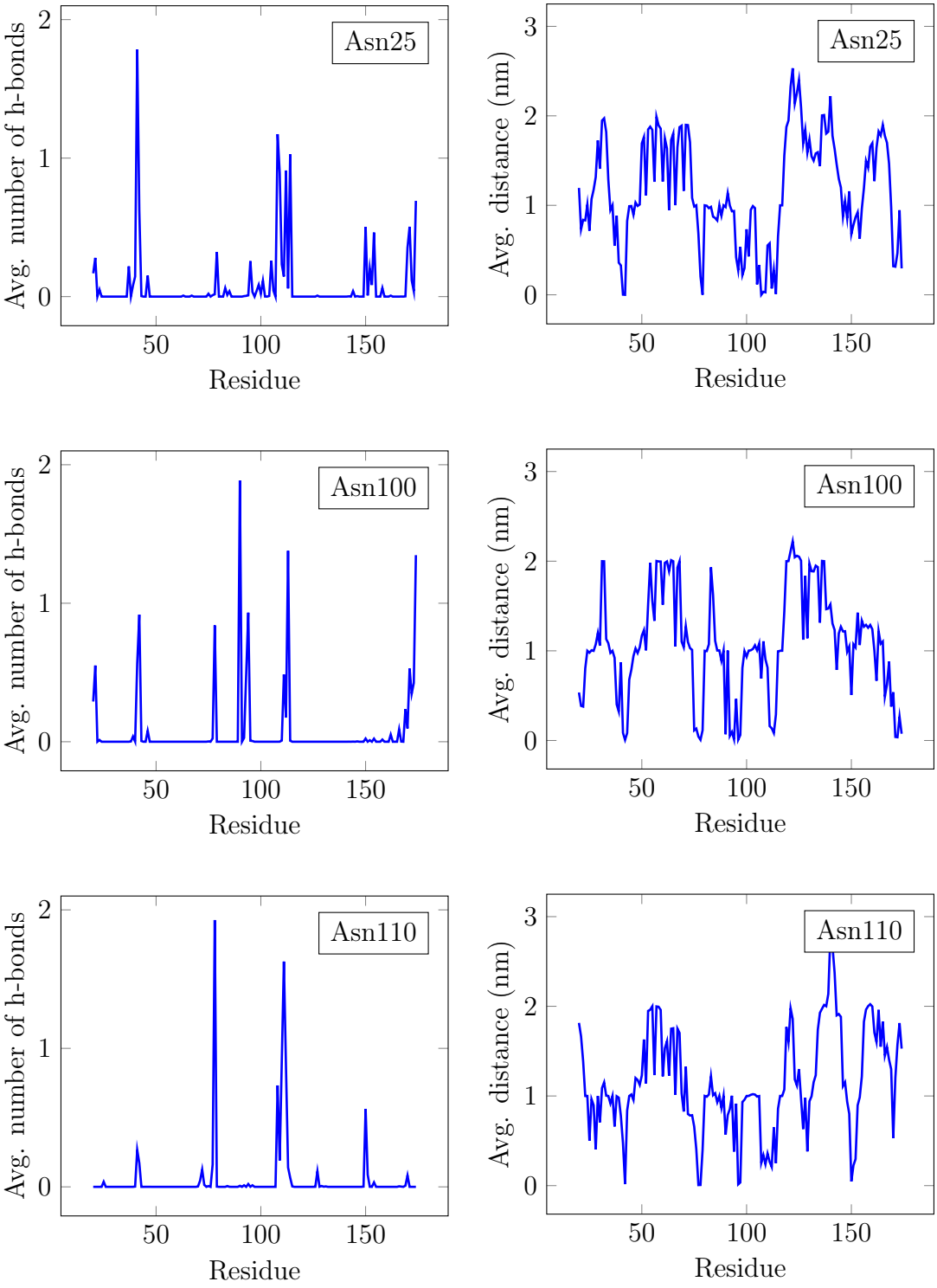


Figure 5.6: Hydrogen bond and minimum distance profiles for the glycosylated CD44 proteins from the simulations S_9, S_{10}, and S_{11}, and the legends give the name of the linking residue. The first 50 ns were not included in the analysis to make sure that the profiles represent real binding interactions. Furthermore, the residues 170 to 174 in the profiles now represent the carbohydrate units of the attached pentasaccharide.

glycosylations seem to have no effect at all. In all cases the overall shape of the hydrogen bond and minimum distance profiles correlates with the non-glycosylated ones, displaying only few exceptions. Firstly, of the three simulation trajectories analyzed, only in the simulation S_{10} the bound HA_{16} is interacting with the attached pentasaccharide with more than one hydrogen bond. Interestingly, these interactions are enough to draw the HA strand into the upper part of the binding groove so that it forms on average 3.5 hydrogen bonds with the residues Arg90, Pro93, and Asn94 from which Arg90 has the strongest contribution of 1.88 hydrogen bonds. Secondly, it seems that the HA_{16} ligand is continuously connected to the main binding cluster formed by Arg41 and Arg78 with two hydrogen bonds, but to which one depends on its orientation inside the binding groove. We return to this matter later.

Our results fortify the idea of covalently attached carbohydrates, especially when linked to Asn25, regulating the binding affinity between HA and CD44. Furthermore, in light of our simulation data, all the glycosylation sites are not equally important, as Asn25 blocks the adhesion, Asn100 enhances it, and Asn110 seems to have no apparent effect.

5.3 Conformation Change in the Binding Site Detaches the Ligand

In order to find out which regions of CD44 HABD have the most prominent structural displacements, we analyzed the root-mean-square fluctuations (RMSFs) of the protein backbone from the simulations S_1, S_2, S_3, and S_4. The procedure started by aligning the atomic coordinates of each simulation frame with a reference frame (the end structure of the protein in a given simulation) and continued by computing the mean squared deviation between these coordinates and the ones from the reference frame via equation:

$$\mathrm{RMSF}_i = \sqrt{\frac{1}{t}\sum_{t_j=1}^{t}(\mathbf{r}_i(t_j) - \mathbf{r}_i^{\mathrm{ref}})^2}, \tag{5.2}$$

where $\mathbf{r}_i(t_j)$ is the position of a given atom i, $\mathbf{r}_i^{\mathrm{ref}}$ is the atomic positions in the reference frame and t is the time [161]. We then calculated the residual RMSF profile by averaging the values of individual backbone atoms for each amino acid residue. As seen from Figure 5.7 the results are remarkably homogeneous. The Link module (residues 23–132) is an especially rigid structure with RMSF of only 0.1–0.2 nm. The flanking regions, on the other hand, are more mobile and especially the C-terminal residues show fluctuations of more than 0.6 nm. Previous NMR and x-ray crystallography studies of human CD44 HABD consisting of residues 20 to 178 observed high flexibility particularly for residues 170 to 178 [15]. Therefore, the result obtained here with the residue 20 to 169 protein construct are well justified.

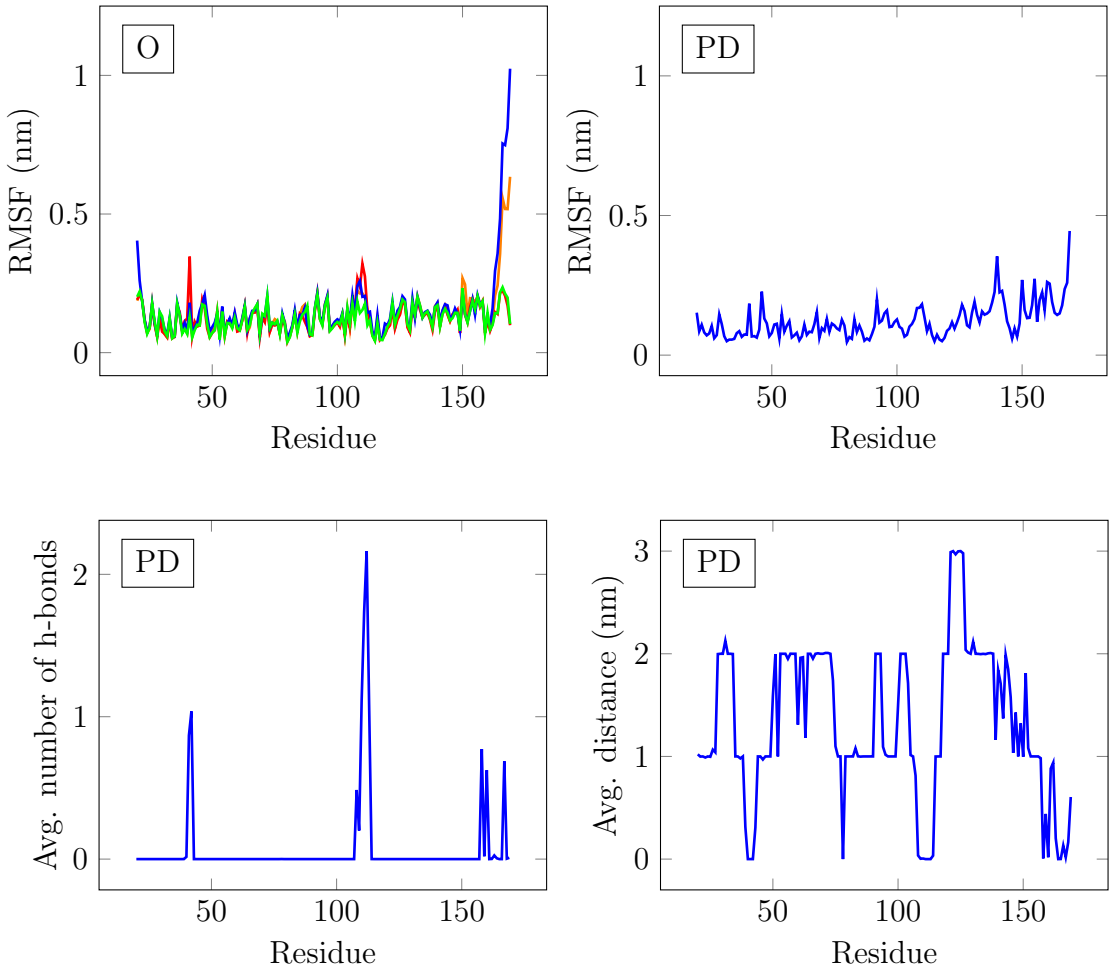


Figure 5.7: The upper left plot gives the RMSF profile of the protein in simulations S_1 (—), S_2 (—), S_3 (—), and S_4 (—). The upper right plot represents the RMS fluctuations of the disordered protein in the simulation S_{17}. The lower row gives the hydrogen bond and minimum distance profiles for the CD44-HA interaction in the simulation S_{18}. The conformational states of the C-terminal tails are given in the legends.

RMSDs were also analyzed as a function of time from simulations S_1, S_2, S_3, and S_4 and they support the above observations of a stable protein (data not shown).

To further assess the significance of the apparent instability in the C-terminus, we induced the PD conformation, where the end residues are freed from the knot-like O-conformation, by manually displacing the residues 150 to 169 from their original coordinates, as shown in the right hand side of Figure 5.8. This was done in two different systems: CD44 without a ligand (simulation S_{17}) and CD44 complexed with HA_8 (simulation S_{18}). The RMSF profile (Figure 5.7) demonstrates that the stability of the protein was not markedly affected upon the O-to-PD transition. While the C-terminal residues from 141 to 169 gain conformational flexibility and experience the unfolding of the β9 strand, the Link module remains unaffected by these perturbations.

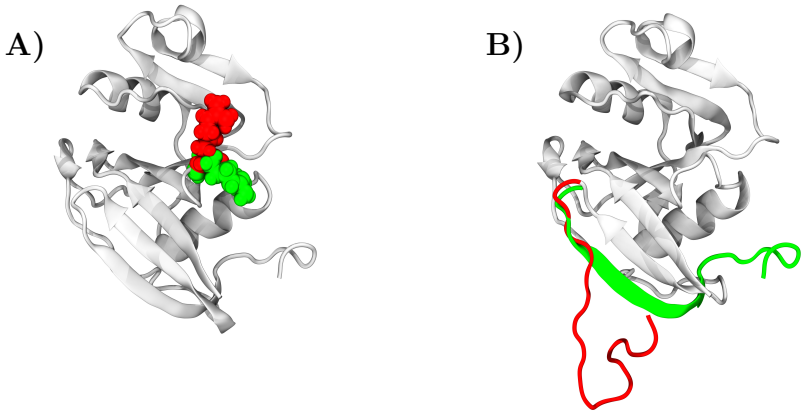


Figure 5.8: Two characteristic conformation changes in CD44 HABD. A) shows the local flipping of the Arg41 side-chain, so that red represents the B-form and green is the A-form. Similarly, B) describes the more profound relaxation of the C-terminal tail. The red is the PD conformation, whereas the green is the crystallographic O-form.

During the course of the simulation S_{18}, the initially disordered C-terminal tail binds the HA_8 oligomer complexed with CD44 at 90 ns. This interaction is then maintained through the rest of the 500 ns simulation, implying that the reached binding conformation was stable. Indeed, the hydrogen bond and the minimum distance profiles in Figure 5.7 demonstrate that the basic residues of the C-terminal end spontaneously bind to the ligand. The C-terminal amino acids most readily interacting with the HA are Lys158, Glu160, and Asp167.

Recently, Favreau et al. [18] acquired the same results with a corresponding system of human CD44 HABD with HA_8 and came to the conclusion that freeing the basic amino acids of the C-terminus will result in high affinity binding. We note that while the PD conformation obviously enhances the binding of HA octamer, it is worthwhile to compare these findings to the binding of HA_{16} to a crystallographic O-state conformation. Consequently, HA_{16} seems to be long enough to bind these basic residues even when the protein is in the O-state (Figure 5.1), thus suggesting that the O-to-PD transition would only matter for very short ($< HA_{16}$) ligands. Our results thereby support the early findings by Peach et al. [13] fortifying the role of the basic residues in the C-terminus, but at the same time question the biological importance of the O-to-PD transition.

Another conformational change characteristic to CD44 is the tilting of the Arg41 side chain (Figure 5.8). We observed this spontaneous A-to-B transition in only two occasions. The first one was, rather surprisingly, in the simulation S_7, where no HA was present. Previously, Jamison et al. [17] and Favreau et al. [18] have claimed that the B-form is the high affinity conformation and that it is only stable

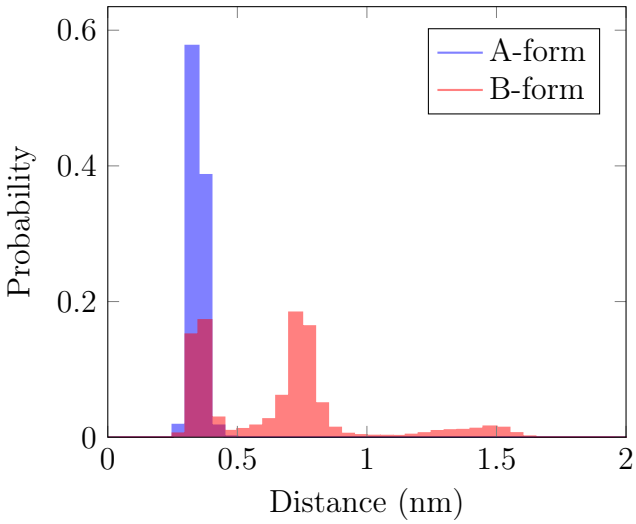


Figure 5.9: The observed minimum distance between the binding site (the Arg78 backbone atoms) and HA_{16} calculated from simulation S_2, so that the first 400 ns represent the A-form and the next 400 ns (400–800 ns into the trajectory) represent the B-form.

when interacting with a strand of HA. Our simulation data, however, clearly suggest that the B-form is stable even in the absence of the HA ligand, because once the transition had commenced, the protein remained at this conformation for the rest of the simulation (a total of 600 ns).

The second A-to-B transition took place in the simulation S_2, which included an HA_8 oligomer docked to the binding groove. Strangely enough, the initially stable CD44-HA interaction was suddenly disconnected after the transition to the B-form occurred at 400-ns mark in the trajectory. In comparison, in the A-form, HA maintained a very narrow distance distribution to the bottom of the binding groove, as plotted in Figure 5.9. Furthermore, after 1600 ns of simulation, the Arg41 side chain spontaneously flipped back to its original A-form orientation, meaning that the protein spent only 1.2 μs in the B-form conformation. This observation is important regarding the regulation of CD44 function, as it suggests that the B-form conformation is actually repelling the bound HA. While these findings contradict heavily with the previous endeavors of deciphering this phenomenon, they are also supported by the fact that in the previous NMR studies of CD44 HABDs in the presence of HA_6, the most prevalent B-form Arg41 side chain rotamers were never recorded [19]. This surprising finding implies that equilibrium conditions might indeed favor the A-form.

A key difference in the A-to-B transition is the tilting of the β1-α1 loop, thus allowing the B-form to commence. To justify our observations, we calculated the characteristic Ramachandran (ψ, ϕ) plots for residues 38 to 45, belonging to this loop. It was done as follows. First we mapped the angle values of each protein backbone ψ, ϕ dihedral pair in every simulation timeframe. The values were then

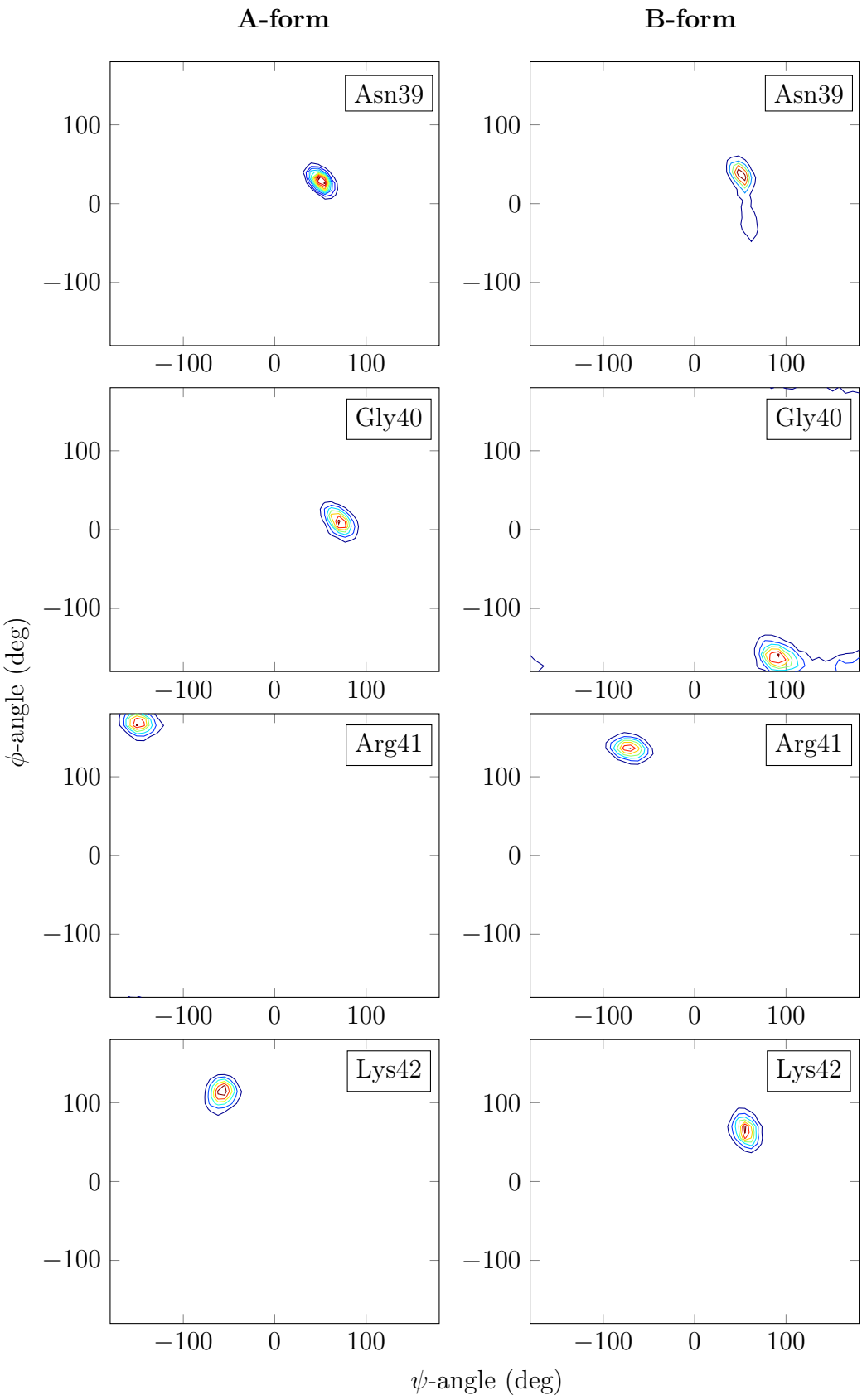


Figure 5.10: Protein backbone (ϕ, ψ) distributions of the loop residues 39 to 42. The values are normalized so that the red contours represent the highest occurrence probabilities.

plotted into an occurrence matrix to obtain the contour plots shown in Figure 5.10. Only the Ramachandran plots of residues 40 to 43 show a considerable change between the A and B-forms. This and the changes itself are consistent with the observations made by Jamison et al. [17], thus verifying that we, indeed, deal with the same phenomenon.

Another characteristic feature of the A-to-B transition is the spatial orientation of the Arg41 side-chain. To measure the changes in this orientation, we calculated the Arg41 side-chain rotamer populations by computing the relative occurence of each possible side-chain dihedral combination before and after the A-to-B transitions from the simulations S_2 and S_7. These rotamer populations are presented in Figure 5.11 and they are clearly more heterogenic (i.e. have wider distributions) in the A-form than in the B-form. In the former, the 12 most probable rotamers account for 85.5 % and 73.3 % of the total sampling when HA is absent or present, respectively, whereas in the latter the same percentages are 97.5 % and 97.9 %. Furthermore, the most probable rotamers in the B-form are not present in the A-form. These observations are again consistent with the ones made by Jamison et al. [17], albeit they have an even narrower distribution representing the B-form. However, their simulations cover timescales of only 20 ns and therefore might not be enough to gather the sampling data that fully represents the underlying phenomena.

As Figure 5.11 shows, the effect of HA's presence is only minor. When it is bound, the rotamer distribution is slightly more narrower, which may suggest that the B-form is more stable in that case or imply that the movements of the Arg41 side-chain are just more constricted in the presence of a ligand.

Based on the whole simulation data, the HA ligand is able to assume two different binding orientations or modes inside the binding groove. Intriguingly, these seem to depend at least to some extent on the conformational state of the protein. For instance, in the A-form by far the most common orientation is the one where the HA oligomer aligns itself over the Arg41 side chain (green strand in Figure 5.12), so that the individual carbohydrate rings orient parallel to the plane formed by the Arg41 side-chain. The second orientation is more exceptional, but it can be induced, for example, by linking a pentasaccharide to Asn100. Now HA is located on the upper part of the binding groove and its carbohydrate rings settle perpendicularly relative to the plane formed by the A-form Arg41 side-chain (red strand in Figure 5.12). Furthermore, in this orientation HA prefers to interact with the Arg78 side chain, whereas in the more common one it connects to Arg41. Interestingly, it seems that only the perpendicular orientation allows the B-form to commence. A reason for this is most likely that in the parallel mode the Arg41 side-chain is forced down by the interacting HA strand, so that the side-chain is unable to flip towards the binding groove. All in all, this finding is important as the orientation of the ligand

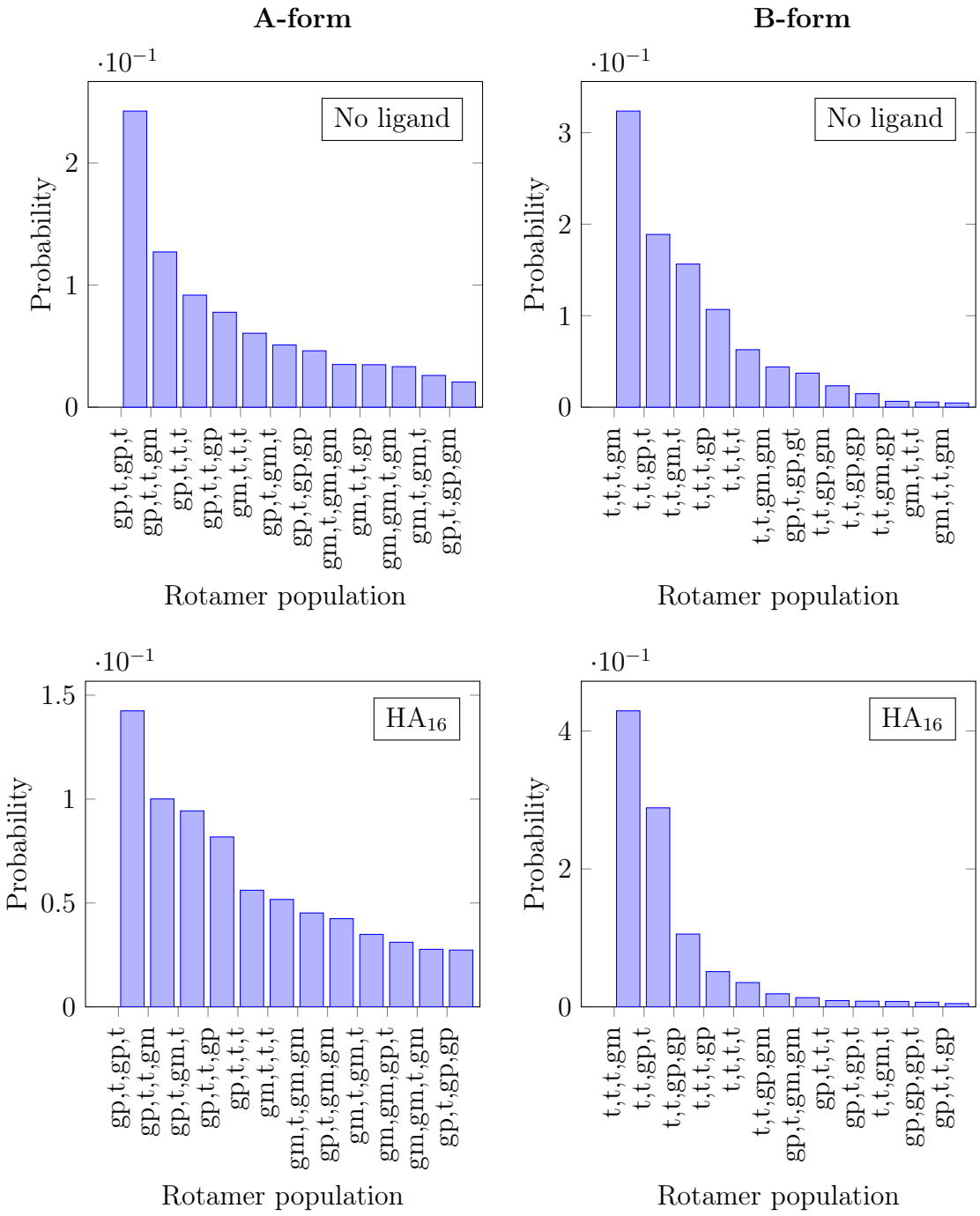


Figure 5.11: Arg41 side-chain rotamer populations in the A-form and the B-form from the simulations S_2 and S_7. Rotamer dihedrals are classified according to the *gauche* conformation, so that *gauche* minus (*gm*) is $-120\,^\circ$ to $0\,^\circ$, *gauche* plus (*gp*) is $0\,^\circ$ to $+120\,^\circ$, and trans (*t*) is $< -120\,^\circ$ to $> +120\,^\circ$.

may play a role in the long-term stability of the binding interaction.

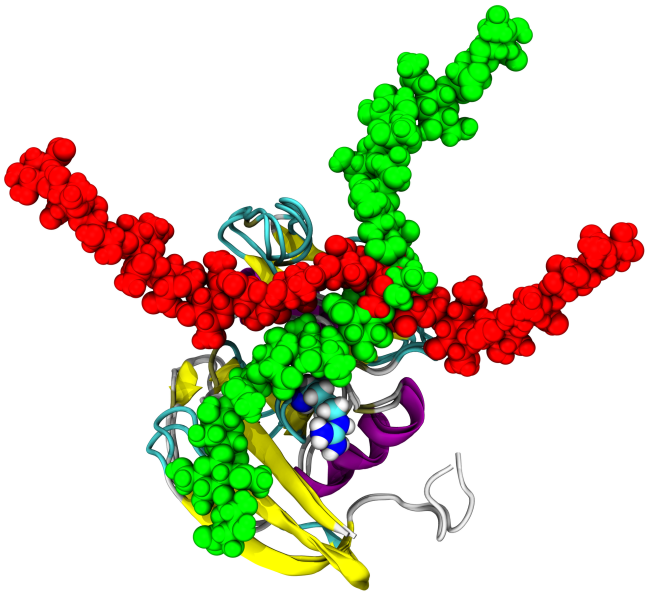


Figure 5.12: Figure illustrating the two possible binding orientation of the HA. The parallel orientation is given in green and the perpendicular orientation in red. Arg41 is also shown for the sake of clarity.

6. CONCLUDING REMARKS

The objective of this Thesis was to study the binding between the CD44 protein and the HA polysaccharide. We employed computational molecular dynamics simulations to explore the role of ligand size, glycosylation, and conformation changes on the regulation of the binding interactions between the two molecules. Importantly, we are the first to implement a model of glycosylated CD44 and thus the first able to investigate these novel areas of CD44 biology. Additionally, unlike previous computational studies, our simulations cover timescales of microseconds, enabling us to have a more comprehensive view into the conformational changes of the protein.

We measured the binding free energy for the HA tetramer and the octamer by using the umbrella sampling method. The result for the former is roughly 7.5 $\mathrm{kJ\,mol^{-1}}$, meaning that it cannot form a stable interaction with the protein. Subsequent analysis of the residual binding profiles reveals that while the principal binding site formed by Arg41 and Arg78 attracts the ligand the most, the residues 108 to 114 from the β5-β6 loop are required to stabilize the interaction. This observation thereby gives a structural explanation to the low-affinity binding of the tetramer, which is too short to interact with these flanking residues. In comparison, the octamer engages the protein with a binding free energy of at least 25 $\mathrm{kJ\,mol^{-1}}$, which corresponds well to an experimentally determined dissociation constant [191]. Given that the clustering of the receptors most likely increases this affinity even more, the obtained value seems rather high in order to reversibly facilitate processes such as the rolling of leukocytes. Therefore, we were left with a question of what mechanism, if any, controls the CD44's binding affinity for longer HA polymers.

CD44 was glycosylated by attaching the core pentasaccharide to each of the residues Asn25, Asn100, and Asn110. Notably, from these the Asn25-linked carbohydrate was the most probable to interfere with the binding of HA, as it assumed a conformation that formed a cover over the principal binding groove. Indeed, a careful analysis of the simulation data and calculation of the free energy profiles for the adsorption of the HA octamer show that the Asn25-linked core pentasaccharide partially masks the native binding site, resulting in a 40 % drop in the binding free energy. Residual binding profiles also confirm the reduced binding affinity. The Asn100-linked core pentasaccharide seems to, on the other hand, strengthen the binding by forming several hydrogen bonds with another free end of the bound lig-

and, whereas the Asn110-linked carbohydrate has no apparent effect. Our results therefore support the idea of site-specific functions of glycosylations.

Previous spectroscopy experiments and computer simulations have observed two distinct conformation transitions in CD44 HABD. The first one is a local tilting of the Arg41 side chain at the main binding site, whereas the second is a much broader relaxation of 19 amino acids in the C-terminal tail. We induced the latter transition manually and found that due to the increased conformational flexibility in the C-terminal end, it is able to bend over to the binding groove and form several direct interactions with the bound ligand. Notably, Favreau et al. [18] made the same observation in a very recent MD simulation study. However, comparing the PD-state residual binding profiles for HA_8 and O-state profiles for HA_{16} shows that the O-to-PD transition is meaningful only for short HA fragments, as HA_{16} is able to bind the C-terminal amino acids already in the O-conformation. Our findings therefore question the biological and physical importance of this transition. In fact, it is somewhat surprising that none of the previous studies have addressed this issue, because it is doubtful that the HABD would assume the knot-like O-conformation through normal folding, given that the C-terminal end is the part that connects to the extracellular stem region in an intact protein. Speculatively, it is possible that expressing HABD as an individual soluble component, without the rest of the protein, results in the protein misfolding to the O-form. Moreover, the O-form might also be a bias caused by the crystallization process, as the available NMR structures are in the flexible PD-form [19]. Nevertheless, these ideas remain to be clarified in future studies.

We also observed the more local A-to-B transition in two simulation trajectories. Most of the previous studies have correlated the "closed" B-form to define a high-affinity state for the binding of HA. Oddly enough, we observed the B-form to repel the otherwise stably bound HA, and furthermore that the presence of a bound ligand is not even a necessity for the transition to occur. Now, combining the above result with the previously observed strong binding affinities, it is tempting to speculate that the B-form is actually an intrinsic molecular mechanism of regulating the CD44-HA interaction by detaching the docked ligand at certain intervals. Yet, the detailed molecular description of this B-form-induced detaching is missing. One hypothesis could be that the colliding Arg41 and Arg78 side chains in the binding site would disrupt the hydrogen bond network with the ligand and drive it to detach. These ideas, however, still require further investigation.

All in all, we found two possible mechanisms, a glycosylation in Asn25 and a local conformation change in the principal binding site, to regulate the otherwise somewhat strong binding affinity. From these, glycosylation is probably a more long-term alternative for this task, as it requires enzymes specialized to covalently linking

the special carbohydrates. Most likely the glycosylation itself is carefully regulated *in situ* at the endoplasmic reticulum or Golgi complex to allow the cell to maintain a specific receptor population at the cell surface. The A-to-B conformation transition, on the other hand, seems to act as a in-built control mechanism. Without this kind of molecular parking meter the interaction with HA, facilitated by receptor clustering, would be too strong and thus unable to support normal migratory functions.

Similar observations have not been made by the previous computational studies. Therefore, we recognize that our results are highly controversial, yet they are also well justified. The previous studies use both mouse and human CD44 HABDs when claiming the B-form is the high-affinity state, meaning that it is not a probable reason for the conflicting findings. In fact, it seems that the only apparent difference in the simulation platform is that both Jamison et al. [17] and Favreau et al. [18] were employing a CHARMM [192] force field. Although the choice of the force field should not affect the results, it cannot be totally ruled out either. More likely the explanation lies in the simulated timescales and the applied restraints in the previous studies. In other words, Jamison et al. [17] used only 20 ns simulation windows to sample the free energy of the A-to-B transition, whereas Favreau et al. [18] restrained the HA oligomer to the binding groove when sampling the same process.

Although this Thesis shows how the binding of HA to its CD44 receptor is regulated and is able to explore novel areas of CD44 biology, it also leaves some fundamental questions open. First and foremost, why does the B-form repel the ligand? Second, can there be an interconversion between the parallel and perpendicular orientations of the bound HA chain to allow the B-form to commence?

We plan to answer these questions by controlling the orientation of the Arg41 side chain with steered MD methods. Furthermore, to fully explore the role of the O-to-PD conformation transition and to eventually have a viable model of the glycocalyx, we need to have the structure of the whole protein, or at least a representative analog, embedded in a lipid bilayer mimicking the physiological composition of the cell membrane. We can also improve our description of the glycosylations to be even more realistic by adding terminal sialic acids to the distal mannoses of each attached core pentasaccharide. Indeed, while this study provides interesting and novel insights into the regulation of protein-carbohydrate interactions, the true work to unravel the sugar puzzle has only just begun.

BIBLIOGRAPHY

[1] **Lipowsky, H.H.**, The endothelial glycocalyx as a barrier to leukocyte adhesion and its mediation by extracellular proteases. *Annals of Biomedical Engineering*, 40(4): 840–848. 2012.

[2] **Evanko, S.P., Tammi, M.I., Tammi, R.H., and Wight, T.N.**, Hyaluronan-dependent pericellular matrix. *Advanced Drug Delivery Reviews*, 59(13): 1351–1365. 2007.

[3] **Weinbaum, S., Tarbell, J.M., and Damiano, E.R.**, The Structure and Function of the Endothelial Glycocalyx Layer. *Annual Review of Biomedical Engineering*, 9(1): 121–67. 2007.

[4] **Reitsma, S., Slaaf, D.W., Vink, H., van Zandvoort, M.A., and oude Egbrink, M.G.**, The endothelial glycocalyx: composition, functions, and visualization. *Pflügers Archiv-European Journal of Physiology*, 454(3): 345–359. 2007.

[5] **Fu, B.M. and Tarbell, J.M.**, Mechano-sensing and transduction by endothelial surface glycocalyx: composition, structure, and function. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*. 2013.

[6] **Naor, D., Sionov, R.V., and Ish-Shalom, D.**, CD44: structure, function and association with the malignant process. *Advances in Cancer Research*, 71: 241–319. 1997.

[7] **Toole, B.P.**, Hyaluronan: from extracellular glue to pericellular cue. *Nature Reviews Cancer*, 4(7): 528–539. 2004.

[8] **Ponta, H., Sherman, L., and Herrlich, P.A.**, CD44: from adhesion molecules to signalling regulators. *Nature Reviews Molecular Cell Biology*, 4(1): 33–45. 2003.

[9] **Lesley, J., English, N., Perschl, A., Gregoroff, J., and Hyman, R.**, Variant cell lines selected for alterations in the function of the hyaluronan receptor CD44 show differences in glycosylation. *The Journal of Experimental Medicine*, 182(2): 431–437. 1995.

[10] **English, N.M., Lesley, J.F., and Hyman, R.**, Site-specific de-N-glycosylation of CD44 can activate hyaluronan binding, and CD44 activation states show distinct threshold densities for hyaluronan binding. *Cancer Research*, 58(16): 3736–3742. 1998.

[11] **Naor, D., Wallach-Dayan, S.B., Zahalka, M.A., and Sionov, R.V.**, Involvement of CD44, a molecule with a thousand faces, in cancer dissemination. In *Seminars in Cancer Biology*, volume 18, pp. 260–267. 2008.

[12] **Zöller, M.**, CD44: can a cancer-initiating cell profit from an abundantly expressed molecule? *Nature Reviews Cancer*, 11(4): 254–267. 2011.

[13] **Peach, R.J., Hollenbaugh, D., Stamenkovic, I., and Aruffo, A.**, Identification of hyaluronic acid binding sites in the extracellular domain of CD44. *The Journal of Cell Biology*, 122(1): 257–264. 1993.

[14] **Bajorath, J., Greenfield, B., Munro, S.B., Day, A.J., and Aruffo, A.**, Identification of CD44 residues important for hyaluronan binding and delineation of the binding site. *Journal of Biological Chemistry*, 273(1): 338–343. 1998.

[15] **Teriete, P., Banerji, S., Noble, M., Blundell, C.D., Wright, A.J., Pickford, A.R., Lowe, E., Mahoney, D.J., Tammi, M.I., and Kahmann, J.D.**, Structure of the regulatory hyaluronan binding domain in the inflammatory leukocyte homing receptor CD44. *Molecular Cell*, 13(4): 483–496. 2004.

[16] **Banerji, S., Wright, A.J., Noble, M., Mahoney, D.J., Campbell, I.D., Day, A.J., and Jackson, D.G.**, Structures of the Cd44–hyaluronan complex provide insight into a fundamental carbohydrate-protein interaction. *Nature Structural & Molecular Biology*, 14(3): 234–239. 2007.

[17] **Jamison II, F.W., Foster, T.J., Barker, J.A., Hills Jr, R.D., and Guvench, O.**, Mechanism of binding site conformational switching in the CD44–hyaluronan protein–carbohydrate binding interaction. *Journal of Molecular Biology*, 406(4): 631–647. 2011.

[18] **Favreau, A.J., Faller, C.E., and Guvench, O.**, CD44 receptor unfolding enhances binding by freeing basic amino acids to contact carbohydrate ligand. *Biophysical Journal*, 105(5): 1217–1226. 2013.

[19] **Takeda, M., Ogino, S., Umemoto, R., Sakakura, M., Kajiwara, M., Sugahara, K.N., Hayasaka, H., Miyasaka, M., Terasawa, H., and Shimada, I.**, Ligand-induced structural changes of the CD44 hyaluronan-binding domain revealed by NMR. *Journal of Biological Chemistry*, 281(52): 40089–40095. 2006.

[20] **Day, A.J. and Prestwich, G.D.**, Hyaluronan-binding proteins: tying up the giant. *Journal of Biological Chemistry*, 277(7): 4585–4588. 2002.

[21] **van Oosten, A.S. and Janmey, P.A.**, Extremely charged and incredibly soft: physical characterization of the pericellular matrix. *Biophysical Journal*, 104(5): 961. 2013.

[22] **Heermann, D.W.**, *Computer-Simulation Methods*. Springer, New York, USA. 1990.

[23] **Bennett, H.S., Luft, J.H., and Hampton, J.C.**, Morphological classifications of vertebrate blood capillaries. *American Journal of Physiology–Legacy Content*, 196(2): 381–390. 1959.

[24] **Curry, F. and Adamson, R.**, Endothelial glycocalyx: permeability barrier and mechanosensor. *Annals of Biomedical Engineering*, 40(4): 828–839. 2012.

[25] **Lee, J.Y. and Spicer, A.P.**, Hyaluronan: a multifunctional, megaDalton, stealth molecule. *Current Opinion in Cell Biology*, 12(5): 581–586. 2000.

[26] **Nelson, D.L., Lehninger, A.L., and Cox, M.M.**, *Lehninger principles of biochemistry*. W.H. Freeman and Company, New York, USA. 2008.

[27] **Lu, T.Y., Lin, B., Kim, J., Sullivan, M., Tobita, K., Salama, G., and Yang, L.**, Repopulation of decellularized mouse heart with human induced pluripotent stem cell-derived cardiovascular progenitor cells. *Nature Communications*, 4. 2013.

[28] **Aubin, H., Kranz, A., Hülsmann, J., Lichtenberg, A., and Akhyari, P.**, Decellularized whole heart for bioartificial heart. In *Cellular Cardiomyoplasty*, pp. 163–178, Springer. 2013.

[29] **Alberts, B., Johnson, A., Lewis, J., Raff, M., and Roberts, K.**, *Molecular Biology of the Cell 4th edition*. Garland Science, New York, USA. 2002.

[30] **VanTeeffelen, J.W., Brands, J., Stroes, E.S., and Vink, H.**, Endothelial glycocalyx: sweet shield of blood vessels. *Trends in Cardiovascular Medicine*, 17(3): 101–105. 2007.

[31] **Schlick, T.**, *Molecular modeling and simulation: an interdisciplinary guide*, volume 21. Springer, New York, USA. 2010.

[32] **Vance, J.E. and Vance, D.E.**, *Biochemistry of lipids, lipoproteins and membranes*. Elsevier, Amsterdam, Netherlands. 2008.

[33] **Rao, V.R., Qasba, P., Balaji, P., and Chandrasekaran, R.**, *Conformation of carbohydrates*. Harwood Academic Publishers, Amsterdam, Netherlands. 1998.

[34] **Almond, A. and Sheehan, J.K.**, Glycosaminoglycan conformation: do aqueous molecular dynamics simulations agree with x-ray fiber diffraction? *Glycobiology*, 10(3): 329–338. 2000.

[35] **Humphrey, W., Dalke, A., and Schulten, K.**, VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1): 33–38. 1996.

[36] **Lipowsky, H.H.**, Microvascular rheology and hemodynamics. *Microcirculation*, 12(1): 5–15. 2005.

[37] **Tarbell, J.M. and Pahakis, M.**, Mechanotransduction and the glycocalyx. *Journal of Internal Medicine*, 259(4): 339–350. 2006.

[38] **Fransson, L.Å., Belting, M., Cheng, F., Jönsson, M., Mani, K., and Sandgren, S.**, Novel aspects of glypican glycobiology. *Cellular and Molecular Life Sciences CMLS*, 61(9): 1016–1024. 2004.

[39] **Carey, D.**, Syndecans: multifunctional cell-surface co-receptors. *Biochemical Journal*, 327(1): 1. 1997.

[40] **Iozzo, R.V.**, Perlecan: a gem of a proteoglycan. *Matrix Biology*, 14(3): 203–208. 1994.

[41] **Kinsella, M.G., Bressler, S.L., and Wight, T.N.**, The regulated synthesis of versican, decorin, and biglycan: extracellular matrix proteoglycans that influence cellular phenotype. *Critical Reviews in Eukaryotic Gene Expression*, 14(3): 203–234. 2004.

[42] **Rapraeger, A.**, Transforming growth factor (type beta) promotes the addition of chondroitin sulfate chains to the cell surface proteoglycan (syndecan) of mouse mammary epithelia. *The Journal of Cell Biology*, 109(5): 2509–2518. 1989.

[43] **Laurent, T.C., Laurent, U.B., and Fraser, J.R.E.**, The structure and function of hyaluronan: an overview. *Immunology and Cell Biology*, 74(2): A1–A7. 1996.

[44] **Stern, R., Asari, A.A., and Sugahara, K.N.**, Hyaluronan fragments: an information-rich system. *European Journal of Cell Biology*, 85(8): 699–715. 2006.

[45] **Kogan, G., Šoltés, L., Stern, R., and Gemeiner, P.**, Hyaluronic acid: a natural biopolymer with a broad range of biomedical and industrial applications. *Biotechnology letters*, 29(1): 17–25. 2007.

[46] **Collins, M.N. and Birkinshaw, C.**, Hyaluronic Acid based Scaffolds for Tissue Engineering–A Review. *Carbohydrate Polymers*, 92(2): 1262–1279. 2013.

[47] **Yoon, H.Y., Koo, H., Choi, K.Y., Chan Kwon, I., Choi, K., Park, J.H., and Kim, K.**, Photo-crosslinked hyaluronic acid nanoparticles with improved stability for *in vivo* tumor-targeted drug delivery. *Biomaterials*, 34(21): 5273–5280. 2013.

[48] **Dahl, L., Dahl, I., Engström-Laurent, A., and Granath, K.**, Concentration and molecular weight of sodium hyaluronate in synovial fluid from patients with rheumatoid arthritis and other arthropathies. *Annals of the Rheumatic Diseases*, 44(12): 817–822. 1985.

[49] **Sattar, A., Kumar, S., and West, D.C.**, Does hyaluronan have a role in endothelial cell proliferation of the synovium? *Seminars in Arthritis and Rheumatism*, 22(1): 37–43. 1992.

[50] **Itano, N., Sawai, T., Yoshida, M., Lenas, P., Yamada, Y., Imagawa, M., Shinomura, T., Hamaguchi, M., Yoshida, Y., and Ohnuki, Y.**, Three isoforms of mammalian hyaluronan synthases have distinct enzymatic properties. *Journal of Biological Chemistry*, 274(35): 25085–25092. 1999.

[51] **Brinck, J. and Heldin, P.**, Expression of recombinant hyaluronan synthase (HAS) isoforms in CHO cells reduces cell migration and cell surface CD44. *Experimental Cell Research*, 252(2): 342–351. 1999.

[52] **Girish, K. and Kemparaju, K.**, The magic glue hyaluronan and its eraser hyaluronidase: a biological overview. *Life Sciences*, 80(21): 1921–1943. 2007.

[53] **Evanko, S.P., Angello, J.C., and Wight, T.N.**, Formation of hyaluronan-and versican-rich pericellular matrix is required for proliferation and migration of vascular smooth muscle cells. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 19(4): 1004–1013. 1999.

[54] **Morawski, M., Brückner, G., Arendt, T., and Matthews, R.**, Aggrecan: Beyond cartilage and into the brain. *The International Journal of Biochemistry & Cell Biology*, 44(5): 690–693. 2012.

[55] **Bono, P., Rubin, K., Higgins, J.M., and Hynes, R.O.**, Layilin, a novel integral membrane protein, is a hyaluronan receptor. *Molecular Biology of the Cell*, 12(4): 891–900. 2001.

[56] **Banerji, S., Ni, J., Wang, S.X., Clasper, S., Su, J., Tammi, R., Jones, M., and Jackson, D.G.**, LYVE-1, a new homologue of the CD44 glycoprotein, is a lymph-specific receptor for hyaluronan. *The Journal of Cell Biology*, 144(4): 789–801. 1999.

[57] **Tsifrina, E., Ananyeva, N.M., Hastings, G., and Liau, G.**, Identification and characterization of three cDNAs that encode putative novel hyaluronan-binding proteins, including an endothelial cell-specific hyaluronan receptor. *The American Journal of Pathology*, 155(5): 1625–1633. 1999.

[58] **Nandi, A., Estess, P., and Siegelman, M.H.**, Hyaluronan anchoring and regulation on the surface of vascular endothelial cells is mediated through the functionally active form of CD44. *Journal of Biological Chemistry*, 275(20): 14939–14948. 2000.

[59] **Nedvetzki, S., Gonen, E., Assayag, N., Reich, R., Williams, R.O., Thurmond, R.L., Huang, J.F., Neudecker, B.A., Wang, F.S., and Turley, E.A.**, RHAMM, a receptor for hyaluronan-mediated motility, compensates for CD44 in inflamed CD44-knockout mice: a different interpretation of redundancy. *Proceedings of the National Academy of Sciences*, 101(52): 18081–18086. 2004.

[60] **Franzén, A., Björnsson, S., and Heinegård, D.**, Cartilage proteoglycan aggregate formation. Role of link protein. *The Biochemical Journal*, 197(3): 669–674. 1981.

[61] **Rauch, U., Karthikeyan, L., Maurel, P., Margolis, R.U., and Margolis, R.K.**, Cloning and primary structure of neurocan, a developmentally regulated, aggregating chondroitin sulfate proteoglycan of brain. *Journal of Biological Chemistry*, 267(27): 19536–19547. 1992.

[62] **Yamada, H., Watanabe, K., Shimonaka, M., and Yamaguchi, Y.**, Molecular cloning of brevican, a novel brain proteoglycan of the aggrecan/versican family. *Journal of Biological Chemistry*, 269(13): 10119–10126. 1994.

[63] **Lesley, J., Gál, I., Mahoney, D.J., Cordell, M.R., Rugg, M.S., Hyman, R., Day, A.J., and Mikecz, K.**, TSG-6 modulates the interaction between hyaluronan and cell surface CD44. *Journal of Biological Chemistry*, 279(24): 25745–25754. 2004.

[64] **Pratt, W.B., Silverstein, A.M., and Galigniana, M.D.**, A model for the cytoplasmic trafficking of signalling proteins involving the hsp90-binding immunophilins and p50 *cdc37*. *Cellular Signalling*, 11(12): 839–851. 1999.

[65] **Deb, T.B. and Datta, K.**, Molecular cloning of human fibroblast hyaluronic acid-binding protein confirms its identity with p-32, a protein co-purified with splicing factor SF2 hyaluronic acid-binding protein as p-32 protein, co-purified with splicing factor SF2. *Journal of Biological Chemistry*, 271(4): 2206–2212. 1996.

[66] **Day, A.J. and de la Motte, C.A.**, Hyaluronan cross-linking: a protective mechanism in inflammation? *Trends in Immunology*, 26(12): 637–643. 2005.

[67] **Hascall, V.C., Majors, A.K., de la Motte, C.A., Evanko, S.P., Wang, A., Drazba, J.A., Strong, S.A., and Wight, T.N.**, Intracellular hyaluronan: a new frontier for inflammation? *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1673(1): 3–12. 2004.

[68] **Selbi, W., De la Motte, C., Hascall, V., Day, A., Bowen, T., and Phillips, A.**, Characterization of hyaluronan cable structure and function in renal proximal tubular epithelial cells. *Kidney International*, 70(7): 1287–1295. 2006.

[69] **Majors, A.K., Austin, R.C., Carol, A., Pyeritz, R.E., Hascall, V.C., Kessler, S.P., Sen, G., and Strong, S.A.**, Endoplasmic reticulum stress induces hyaluronan deposition and leukocyte adhesion. *Journal of Biological Chemistry*, 278(47): 47223–47231. 2003.

[70] **Bruegger, D., Jacob, M., Rehm, M., Loetsch, M., Welsch, U., Conzen, P., and Becker, B.F.**, Atrial natriuretic peptide induces shedding of endothelial glycocalyx in coronary vascular bed of guinea pig hearts. *American Journal of Physiology-Heart and Circulatory Physiology*, 289(5): H1993–H1999. 2005.

[71] **Arkill, K., Knupp, C., Michel, C., Neal, C., Qvortrup, K., Rostgaard, J., and Squire, J.**, Similar endothelial glycocalyx structures in microvessels from a range of mammalian tissues: evidence for a common filtering mechanism? *Biophysical Journal*, 101(5): 1046–1056. 2011.

[72] **McLane, L.T., Chang, P., Granqvist, A., Boehm, H., Kramer, A., Scrimgeour, J., and Curtis, J.E.**, Spatial organization and mechanical properties of the pericellular matrix on chondrocytes. *Biophysical Journal*, 104(5): 986–996. 2013.

[73] **Vink, H. and Duling, B.R.**, Capillary endothelial surface layer selectively reduces plasma solute distribution volume. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(1): H285–H289. 2000.

[74] **Damiano, E.**, The effect of the endothelial-cell glycocalyx on the motion of red blood cells through capillaries. *Microvascular Research*, 55(1): 77–91. 1998.

[75] **van Haaren, P.M., VanBavel, E., Vink, H., and Spaan, J.A.**, Localization of the permeability barrier to solutes in isolated arteries by confocal microscopy. *American Journal of Physiology-Heart and Circulatory Physiology*, 285(6): H2848–H2856. 2003.

[76] **van Haaren, P.M., VanBavel, E., Vink, H., and Spaan, J.A.**, Charge modification of the endothelial surface layer modulates the permeability barrier of isolated rat mesenteric small arteries. *American Journal of Physiology-Heart and Circulatory Physiology*, 289(6): H2503–H2507. 2005.

[77] **Espinosa-Marzal, R.M., Fontani, G., Reusch, F.B., Roba, M., Spencer, N.D., and Crockett, R.**, Sugars communicate through water: oriented glycans induce water structuring. *Biophysical Journal*, 104(12): 2686–2694. 2013.

[78] **Arisaka, T., Mitsumata, M., Kawasumi, M., Tohjima, T., Hirose, S., and Yoshida, Y.**, Effects of shear stress on glycosaminoglycan synthesis in vascular endothelial cellsa. *Annals of the New York Academy of Sciences*, 748(1): 543–554. 1994.

[79] **Hoover, R., Folger, R., Haering, W., Ware, B., and Karnovsky, M.**, Adhesion of leukocytes to endothelium: roles of divalent cations, surface charge, chemotactic agents and substrate. *Journal of Cell Science*, 45(1): 73–86. 1980.

[80] **Zarbock, A. and Ley, K.**, Neutrophil adhesion and activation under flow. *Microcirculation*, 16(1): 31–42. 2009.

[81] **Luo, B.H., Carman, C.V., and Springer, T.A.**, Structural basis of integrin regulation and signaling. *Annual Review of Immunology*, 25(1): 619–647. 2007.

[82] **Cohen, M., Kam, Z., Addadi, L., and Geiger, B.**, Dynamic study of the transition from hyaluronan-to integrin-mediated adhesion in chondrocytes. *The EMBO journal*, 25(2): 302–311. 2006.

[83] **Wilkinson, T.S., Bressler, S.L., Evanko, S.P., Braun, K.R., and Wight, T.N.**, Overexpression of hyaluronan synthases alters vascular smooth muscle cell phenotype and promotes monocyte adhesion. *Journal of Cellular Physiology*, 206(2): 378–385. 2006.

[84] **Megens, R., Reitsma, S., Schiffers, P., Hilgers, R., De Mey, J., Slaaf, D., and van Zandvoort, M.**, Two-photon microscopy of vital murine elastic and muscular arteries. *Journal of Vascular Research*, 44(2): 87–98. 2006.

[85] **van den Berg, B.M., Vink, H., and Spaan, J.A.**, The endothelial glycocalyx protects against myocardial edema. *Circulation Research*, 92(6): 592–594. 2003.

[86] **Rilla, K., Lammi, M.J., Sironen, R., Törrönen, K., Luukkonen, M., Hascall, V.C., Midura, R.J., Hyttinen, M., Pelkonen, J., and Tammi, M.**, Changed lamellipodial extension, adhesion plaques and migration in epidermal keratinocytes containing constitutively expressed sense and antisense hyaluronan synthase 2 (Has2) genes. *Journal of Cell Science*, 115(18): 3633–3643. 2002.

[87] **Henry, C.B. and Duling, B.R.**, TNF-α increases entry of macromolecules into luminal endothelial cell glycocalyx. *American Journal of Physiology-Heart and Circulatory Physiology*, 279(6): H2815–H2823. 2000.

[88] **Zuurbier, C.J., Demirci, C., Koeman, A., Vink, H., and Ince, C.**, Short-term hyperglycemia increases endothelial glycocalyx permeability and acutely decreases lineal density of capillaries with flowing red blood cells. *Journal of Applied Physiology*, 99(4): 1471–1476. 2005.

[89] **Hofmann-Kiefer, K., Kemming, G., Chappell, D., Flondor, M., Kisch-Wedel, H., Hauser, A., Pallivathukal, S., Conzen, P., and Rehm, M.**, Serum heparan sulfate levels are elevated in endotoxemia. *European Journal of Medical Research*, 14(12): 526. 2009.

[90] **Constantinescu, A.A., Vink, H., and Spaan, J.A.**, Elevated capillary tube hematocrit reflects degradation of endothelial cell glycocalyx by oxidized LDL. *American Journal of Physiology-Heart and Circulatory Physiology*, 280(3): H1051–H1057. 2001.

[91] **Mulivor, A.W. and Lipowsky, H.H.**, Inflammation-and ischemia-induced shedding of venular glycocalyx. *American Journal of Physiology-Heart and Circulatory Physiology*, 286(5): H1672–H1680. 2004.

[92] **Gouverneur, M., Spaan, J.A., Pannekoek, H., Fontijn, R.D., and Vink, H.**, Fluid shear stress stimulates incorporation of hyaluronan into endothelial cell glycocalyx. *American Journal of Physiology-Heart and Circulatory Physiology*, 290(1): H458–H452. 2006.

[93] **Chappell, D., Hofmann-Kiefer, K., Jacob, M., Rehm, M., Briegel, J., Welsch, U., Conzen, P., and Becker, B.F.**, TNF-α induced shedding of the endothelial glycocalyx is prevented by hydrocortisone and antithrombin. *Basic Research in Cardiology*, 104(1): 78–89. 2009.

[94] **Yamazaki, K., Fukuda, K., Matsukawa, M., Hara, F., Matsushita, T., Yamamoto, N., Yoshida, K., Munakata, H., and Hamanishi, C.**, Cyclic tensile stretch loaded on bovine chondrocytes causes depolymerization of hyaluronan: involvement of reactive oxygen species. *Arthritis & Rheumatism*, 48(11): 3151–3158. 2003.

[95] **Toole, B.P.**, Hyaluronan-CD44 interactions in cancer: paradoxes and possibilities. *Clinical Cancer Research*, 15(24): 7462–7468. 2009.

[96] **Johnson, P. and Ruffell, B.**, CD44 and its role in inflammation and inflammatory diseases. *Inflammation & Allergy-Drug Targets*, 8(3): 208–220. 2009.

[97] **Mummert, M.E.**, Immunologic roles of hyaluronan. *Immunologic Research*, 31(3): 189–205. 2005.

[98] **Dalchau, R., Kirkley, J., and Fabre, J.W.**, Monoclonal antibody to a human brain-granulocyte-T lymphocyte antigen probably homologous to the W 3/13 antigen of the rat. *European Journal of Immunology*, 10(10): 745–749. 1980.

[99] **Günthert, U., Stauder, R., Mayer, B., Terpe, H.J., Finke, L., and Friedrichs, K.**, Are CD44 variant isoforms involved in human tumour progression? *Cancer Surveys*, 24(1): 19–42. 1995.

[100] **Goodfellow, P.N., Banting, G., Wiles, M.V., Tunnacliffe, A., Parkar, M., Solomon, E., Dalchau, R., and Fabre, J.W.**, The gene, MIC4, which controls expression of the antigen defined by monoclonal antibody F10. 44.2, is on human chromosome 11. *European Journal of Immunology*, 12(8): 659–663. 1982.

[101] **Screaton, G.R., Bell, M.V., Jackson, D.G., Cornelis, F.B., Gerth, U., and Bell, J.I.**, Genomic structure of DNA encoding the lymphocyte homing receptor CD44 reveals at least 12 alternatively spliced exons. *Proceedings of the National Academy of Sciences*, 89(24): 12160–12164. 1992.

[102] **Screaton, G., Bell, M., Bell, J., and Jackson, D.**, The identification of a new alternative exon with highly restricted tissue expression in transcripts encoding the mouse Pgp-1 (CD44) homing receptor. Comparison of all 10 variable exons between

mouse, human, and rat. *Journal of Biological Chemistry*, 268(17): 12235–12238. 1993.

[103] **Jijiwa, M., Demir, H., Gupta, S., Leung, C., Joshi, K., Orozco, N., Huang, T., Yildiz, V.O., Shibahara, I., and de Jesus, J.A.**, CD44v6 regulates growth of brain tumor stem cells partially through the AKT-mediated pathway. *PloS One*, 6(9): e24217. 2011.

[104] **Olsson, E., Honeth, G., Bendahl, P.O., Saal, L., Gruvberger-Saal, S., Ringnér, M., Vallon-Christersson, J., Jönsson, G., Holm, K., and Lövgren, K.**, CD44 isoforms are heterogeneously expressed in breast cancer and correlate with tumor subtypes and cancer stem cell markers. *BMC Cancer*, 11(1): 418. 2011.

[105] **Nagano, O., Okazaki, S., and Saya, H.**, Redox regulation in stem-like cancer cells by CD44 variant isoforms. *Oncogene*, 32(44): 5191–5198. 2013.

[106] **Stamenkovic, I., Aruffo, A., Amiot, M., and Seed, B.**, The hematopoietic and epithelial forms of CD44 are distinct polypeptides with different adhesion potentials for hyaluronate-bearing cells. *The EMBO Journal*, 10(2): 343. 1991.

[107] **Greenfield, B., Wang, W.C., Marquardt, H., Piepkorn, M., Wolff, E.A., Aruffo, A., and Bennett, K.L.**, Characterization of the heparan sulfate and chondroitin sulfate assembly sites in CD44. *Journal of Biological Chemistry*, 274(4): 2511–2517. 1999.

[108] **Banerji, S., Day, A.J., Kahmann, J.D., and Jackson, D.G.**, Characterization of a functional hyaluronan-binding domain from the human CD44 molecule expressed in *Escherichia coli*. *Protein Expression and Purification*, 14(3): 371–381. 1998.

[109] **Okamoto, I., Kawano, Y., Murakami, D., Sasayama, T., Araki, N., Miki, T., Wong, A.J., and Saya, H.**, Proteolytic release of CD44 intracellular domain and its role in the CD44 signaling pathway. *The Journal of Cell Biology*, 155(5): 755–762. 2001.

[110] **Okamoto, I., Tsuiki, H., Kenyon, L.C., Godwin, A.K., Emlet, D.R., Holgado-Madruga, M., Lanham, I.S., Joynes, C.J., Vo, K.T., and Guha, A.**, Proteolytic cleavage of the CD44 adhesion molecule in multiple human tumors. *The American Journal of Pathology*, 160(2): 441–447. 2002.

[111] **Kalnina, Z., Zayakin, P., Silina, K., and Linē, A.**, Alterations of pre-mRNA splicing in cancer. *Genes, Chromosomes and Cancer*, 42(4): 342–357. 2005.

[112] **Zhang, L., David, G., and Esko, J.D.**, Repetitive Ser-Gly sequences enhance heparan sulfate assembly in proteoglycans. *Journal of Biological Chemistry*, 270(45): 27127–27135. 1995.

[113] **Tremmel, M., Matzke, A., Albrecht, I., Laib, A.M., Olaku, V., Ballmer-Hofer, K., Christofori, G., Héroult, M., Augustin, H.G., and Ponta, H.**, A CD44v6 peptide reveals a role of CD44 in VEGFR-2 signaling and angiogenesis. *Blood*, 114(25): 5236–5244. 2009.

[114] **Orian-Rousseau, V. and Ponta, H.**, Adhesion proteins meet receptors: a common theme? *Advances in Cancer Research*, 101(1): 63–92. 2008.

[115] **Bennett, K.L., Jackson, D.G., Simon, J.C., Tanczos, E., Peach, R., Modrell, B., Stamenkovic, I., Plowman, G., and Aruffo, A.**, CD44 isoforms containing exon V3 are responsible for the presentation of heparin-binding growth factor. *The Journal of Cell Biology*, 128(4): 687–698. 1995.

[116] **Liu, D. and Sy, M.S.**, Phorbol myristate acetate stimulates the dimerization of CD44 involving a cysteine in the transmembrane domain. *The Journal of Immunology*, 159(6): 2702–2711. 1997.

[117] **Oliferenko, S., Paiha, K., Harder, T., Gerke, V., Schwärzler, C., Schwarz, H., Beug, H., Günthert, U., and Huber, L.A.**, Analysis of Cd44-containing lipid rafts recruitment of annexin II and stabilization by the actin cytoskeleton. *The Journal of Cell Biology*, 146(4): 843–854. 1999.

[118] **Pure, E., Camp, R.L., Peritt, D., Panettieri, R., Lazaar, A.L., and Nayak, S.**, Defective phosphorylation and hyaluronate binding of CD44 with point mutations in the cytoplasmic domain. *The Journal of Experimental Medicine*, 181(1): 55–62. 1995.

[119] **Couchman, J.R.**, Transmembrane signaling proteoglycans. *Annual Review of Cell and Developmental Biology*, 26(1): 89–114. 2010.

[120] **Naor, D., Nedvetzki, S., Golan, I., Melnik, L., and Faitelson, Y.**, CD44 in cancer. *Critical Reviews in Clinical Laboratory Sciences*, 39(6): 527–579. 2002.

[121] **Hanley, W.D., Burdick, M.M., Konstantopoulos, K., and Sackstein, R.**, CD44 on LS174T colon carcinoma cells possesses E-selectin ligand activity. *Cancer Research*, 65(13): 5812–5817. 2005.

[122] **Napier, S.L., Healy, Z.R., Schnaar, R.L., and Konstantopoulos, K.**, Selectin ligand expression regulates the initial vascular interactions of colon carcinoma cells: the roles of CD44V and alternative sialofucosylated selectin ligands. *Journal of Biological Chemistry*, 282(6): 3433–3441. 2007.

[123] **Denhardt, D.T., Noda, M., O'Regan, A.W., Pavlin, D., Berman, J.S., et al.**, Osteopontin as a means to cope with environmental insults: regulation of inflammation, tissue remodeling, and cell survival. *Journal of Clinical Investigation*, 107(9): 1055–1061. 2001.

[124] **Bourguignon, L.Y., Zhu, H., Shao, L., Zhu, D., and Chen, Y.W.**, Rho-kinase (ROK) promotes CD44v3, 8–10-ankyrin interaction and tumor cell migration in metastatic breast cancer cells. *Cell Motility and the Cytoskeleton*, 43(4): 269–287. 1999.

[125] **Bourguignon, L.Y., Zhu, H., Shao, L., and Chen, Y.W.**, CD44 interaction with tiam1 promotes Rac1 signaling and hyaluronic acid-mediated breast tumor cell migration. *Journal of Biological Chemistry*, 275(3): 1829–1838. 2000.

[126] **Bourguignon, L.Y.**, CD44-mediated oncogenic signaling and cytoskeleton activation during mammary tumor progression. *Journal of Mammary Gland Biology and Neoplasia*, 6(3): 287–297. 2001.

[127] **Bourguignon, L.Y., Zhu, H., Zhou, B., Diedrich, F., Singleton, P.A., and Hung, M.C.**, Hyaluronan promotes CD44v3-Vav2 interaction with Grb2-p185HER2 and induces Rac1 and Ras signaling during ovarian tumor cell migration and growth. *Journal of Biological Chemistry*, 276(52): 48679–48692. 2001.

[128] **Oliferenko, S., Kaverina, I., Small, J.V., and Huber, L.A.**, Hyaluronic acid (HA) binding to CD44 activates Rac1 and induces lamellipodia outgrowth. *The Journal of Cell Biology*, 148(6): 1159–1164. 2000.

[129] **Foger, N., Marhaba, R., and Zoller, M.**, Involvement of CD44 in cytoskeleton rearrangement and raft reorganization in T cells. *Journal of Cell Science*, 114(6): 1169–1178. 2001.

[130] **Turley, E.A., Noble, P.W., and Bourguignon, L.Y.**, Signaling properties of hyaluronan receptors. *Journal of Biological Chemistry*, 277(7): 4589–4592. 2002.

[131] **Bourguignon, L.Y., Spevak, C.C., Wong, G., Xia, W., and Gilad, E.**, Hyaluronan-CD44 interaction with protein kinase C promotes oncogenic signaling by the stem cell marker Nanog and the production of microRNA-21, leading to down-regulation of the tumor suppressor protein PDCD4, anti-apoptosis, and chemotherapy resistance in breast tumor cells. *Journal of Biological Chemistry*, 284(39): 26533–26546. 2009.

[132] **Tsatas, D., Kanagasundaram, V., Kaye, A., and Novak, U.**, EGF receptor modifies cellular responses to hyaluronan in glioblastoma cell lines. *Journal of Clinical Neuroscience*, 9(3): 282–288. 2002.

[133] **Tsukita, S. and Yonemura, S.**, Cortical actin organization: lessons from ERM (ezrin/radixin/moesin) proteins. *Journal of Biological Chemistry*, 274(49): 34507–34510. 1999.

[134] **König, H., Ponta, H., and Herrlich, P.**, Coupling of signal transduction to alternative pre-mRNA splicing by a composite splice regulator. *The EMBO Journal*, 17(10): 2904–2913. 1998.

[135] **Morrison, H., Sherman, L.S., Legg, J., Banine, F., Isacke, C., Haipek, C.A., Gutmann, D.H., Ponta, H., and Herrlich, P.**, The NF2 tumor suppressor gene product, merlin, mediates contact inhibition of growth through interactions with CD44. *Genes & Development*, 15(8): 968–980. 2001.

[136] **Zhu, D. and Bourguignon, L.Y.**, Interaction between CD44 and the repeat domain of ankyrin promotes hyaluronic acid-mediated ovarian tumor cell migration. *Journal of Cellular Physiology*, 183(2): 182–195. 2000.

[137] **Bourguignon, L.Y., Singleton, P.A., Zhu, H., and Diedrich, F.**, Hyaluronan-mediated CD44 interaction with RhoGEF and Rho kinase promotes Grb2-associated binder-1 phosphorylation and phosphatidylinositol 3-kinase signaling leading to cytokine (macrophage-colony stimulating factor) production and breast tumor progression. *Journal of Biological Chemistry*, 278(32): 29420–29434. 2003.

[138] **Lokeshwar, V.B., Fregien, N., and Bourguignon, L.**, Ankyrin-binding domain of CD44 (GP85) is required for the expression of hyaluronic acid-mediated adhesion function. *The Journal of Cell Biology*, 126(4): 1099–1109. 1994.

[139] **Ghatak, S., Misra, S., and Toole, B.P.**, Hyaluronan oligosaccharides inhibit anchorage-independent growth of tumor cells by suppressing the phosphoinositide 3-kinase/Akt cell survival pathway. *Journal of Biological Chemistry*, 277(41): 38013–38020. 2002.

[140] **Peterson, R.M., Yu, Q., Stamenkovic, I., and Toole, B.P.**, Perturbation of hyaluronan interactions by soluble CD44 inhibits growth of murine mammary carcinoma cells in ascites. *The American Journal of Pathology*, 156(6): 2159–2167. 2000.

[141] **Ward, J.A., Huang, L., Guo, H., Ghatak, S., and Toole, B.P.**, Perturbation of hyaluronan interactions inhibits malignant properties of glioma cells. *The American Journal of Pathology*, 162(5): 1403–1409. 2003.

[142] **Naor, D., Nedvetzki, S., Walmsley, M., Yayon, A., Turley, E.A., Golan, I., Caspi, D., Sebban, L.E., Zick, Y., Garin, T., et al.**, CD44 involvement in autoimmune inflammations. *Annals of the New York Academy of Sciences*, 1110(1): 233–247. 2007.

[143] **Noble, P.W., McKee, C.M., Cowman, M., and Shin, H.S.**, Hyaluronan fragments activate an NF-kappa B/I-kappa B alpha autoregulatory loop in murine macrophages. *The Journal of Experimental Medicine*, 183(5): 2373–2378. 1996.

[144] **DeGrendele, H.C., Estess, P., Picker, L.J., and Siegelman, M.H.**, CD44 and its ligand hyaluronate mediate rolling under physiologic flow: a novel lymphocyte-endothelial cell primary adhesion pathway. *The Journal of Experimental Medicine*, 183(3): 1119–1130. 1996.

[145] **Siegelman, M.H., Stanescu, D., Estess, P., et al.**, The CD44-initiated pathway of T-cell extravasation uses VLA-4 but not LFA-1 for firm adhesion. *Journal of Clinical Investigation*, 105(5): 683–691. 2000.

[146] **Skelton, T.P., Zeng, C., Nocks, A., and Stamenkovic, I.**, Glycosylation provides both stimulatory and inhibitory effects on cell surface and soluble CD44 binding to hyaluronan. *The Journal of Cell Biology*, 140(2): 431–446. 1998.

[147] **Bartolazzi, A., Jackson, D., Bennett, K., Aruffo, A., Dickinson, R., Shields, J., Whittle, N., and Stamenkovic, I.**, Regulation of growth and dissemination of a human lymphoma by CD44 splice variants. *Journal of Cell Science*, 108(4): 1723–1733. 1995.

[148] **Bennett, K.L., Modrell, B., Greenfield, B., Bartolazzi, A., Stamenkovic, I., Peach, R., Jackson, D.G., Spring, F., and Aruffo, A.**, Regulation of CD44 binding to hyaluronan by glycosylation of variably spliced exons. *The Journal of Cell Biology*, 131(6): 1623–1633. 1995.

[149] **Wolny, P.M., Banerji, S., Gounou, C., Brisson, A.R., Day, A.J., Jackson, D.G., and Richter, R.P.**, Analysis of cd44-hyaluronan interactions in an artificial membrane system insights into the distinct binding properties of high and low molecular weight hyaluronan. *Journal of Biological Chemistry*, 285(39): 30170–30180. 2010.

[150] **Culty, M., Nguyen, H.A., and Underhill, C.B.**, The hyaluronan receptor (CD44) participates in the uptake and degradation of hyaluronan. *The Journal of Cell Biology*, 116(4): 1055–1062. 1992.

[151] **Kothapalli, D., Flowers, J., Xu, T., Puré, E., and Assoian, R.K.**, Differential activation of ERK and Rac mediates the proliferative and anti-proliferative effects of hyaluronan and CD44. *Journal of Biological Chemistry*, 283(46): 31823–31829. 2008.

[152] **Yu, Q., Toole, B.P., and Stamenkovic, I.**, Induction of apoptosis of metastatic mammary carcinoma cells in vivo by disruption of tumor cell surface CD44 function. *The Journal of Experimental Medicine*, 186(12): 1985–1996. 1997.

[153] **Rochman, M., Moll, J., Herrlich, P., Wallach, S.B., Nedvetzki, S., Sionov, R.V., Golan, I., Ish-Shalom, D., and Naor, D.**, The CD44 receptor of lymphoma cells: structure-function relationships and mechanism of activation. *Cell Communication and Adhesion*, 7(4): 331–347. 2000.

[154] **Lesley, J., Hyman, R., English, N., Catterall, J.B., and Turner, G.A.**, CD44 in inflammation and metastasis. *Glycoconjugate Journal*, 14(5): 611–622. 1997.

[155] **Katoh, S., Zheng, Z., Oritani, K., Shimozato, T., and Kincade, P.W.**, Glycosylation of CD44 negatively regulates its recognition of hyaluronan. *The Journal of Experimental Medicine*, 182(2): 419–429. 1995.

[156] **Takeda, M., Terasawa, H., Sakakura, M., Yamaguchi, Y., Kajiwara, M., Kawashima, H., Miyasaka, M., and Shimada, I.**, Hyaluronan recognition mode of CD44 revealed by cross-saturation and chemical shift perturbation experiments. *Journal of Biological Chemistry*, 278(44): 43550–43555. 2003.

[157] **Takeda, M., Terasawa, H., Sakakura, M., Yamaguchi, Y., Kajiwara, M., Kawashima, H., Miyasaka, M., and Shimada, I.**, Letter to the Editor: 1 H, 13 C and 15 N backbone resonance assignments of the hyaluronan-binding domain of CD44. *Journal of Biomolecular NMR*, 29(1): 97–98. 2004.

[158] **Ogino, S., Nishida, N., Umemoto, R., Suzuki, M., Takeda, M., Terasawa, H., Kitayama, J., Matsumoto, M., Hayasaka, H., and Miyasaka, M.**, Two-state conformations in the hyaluronan-binding domain regulate CD44 adhesiveness under flow condition. *Structure*, 18(5): 649–656. 2010.

[159] **Allen, M.P. and Tildesley, D.J.**, *Computer simulation of liquids*. Oxford university press, New York, USA. 1989.

[160] **Alder, B.J. and Wainwright, T.**, Studies in molecular dynamics. I. General method. *The Journal of Chemical Physics*, 31(2): 459–466. 1959.

[161] **van der Spoel, D., Lindahl, E., Hess, B., Van Buuren, A., Apol, E., Meulenhoff, P., Tieleman, D., Sijbers, A., Feenstra, K., and van Drunen, R.**, Gromacs User manual version 4.6.3. 2013.

[162] **Orsi, M. and Essex, J.W.**, Permeability of drugs and hormones through a lipid bilayer: insights from dual-resolution molecular dynamics. *Soft Matter*, 6(16): 3797–3808. 2010.

[163] **Falck, E., Róg, T., Karttunen, M., and Vattulainen, I.**, Lateral diffusion in lipid membranes through collective flows. *Journal of the American Chemical Society*, 130(1): 44–45. 2008.

[164] **van den Bogaart, G., Meyenberg, K., Risselada, H.J., Amin, H., Willig, K.I., Hubrich, B.E., Dier, M., Hell, S.W., Grubmüller, H., Diederichsen, U., et al.**, Membrane protein sequestering by ionic protein-lipid interactions. *Nature*, 479(7374): 552–555. 2011.

[165] **Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A.**, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19): 5179–5197. 1995.

[166] **Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O., and Shaw, D.E.**, Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8): 1950–1958. 2010.

[167] **Kirschner, K.N., Yongye, A.B., Tschampel, S.M., González-Outeiriño, J., Daniels, C.R., Foley, B.L., and Woods, R.J.**, GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *Journal of Computational Chemistry*, 29(4): 622–655. 2008.

[168] **Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., and Berendsen, H.J.**, GROMACS: fast, flexible, and free. *Journal of Computational Chemistry*, 26(16): 1701–1718. 2005.

[169] **Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E.**, GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3): 435–447. 2008.

[170] **Van Gunsteren, W. and Berendsen, H.**, A leap-frog algorithm for stochastic dynamics. *Molecular Simulation*, 1(3): 173–185. 1988.

[171] **Darden, T., York, D., and Pedersen, L.**, Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12): 10089–10092. 1993.

[172] **Essmann, U., Perera, L., Berkowitz, M.L., Darden, T., Lee, H., and Pedersen, L.G.**, A smooth particle mesh Ewald method. *Journal of Chemical Physics*, 103(19): 8577–8593. 1995.

[173] **Ewald, P.P.**, Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik*, 369(3): 253–287. 1921.

[174] **Bussi, G., Donadio, D., and Parrinello, M.**, Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1): 4101. 2007.

[175] **Parrinello, M. and Rahman, A.**, Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12): 7182–7190. 1981.

[176] **Hess, B., Bekker, H., Berendsen, H.J., and Fraaije, J.G.**, LINCS: a linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12): 1463–1472. 1997.

[177] **Miyamoto, S. and Kollman, P.A.**, SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8): 952–962. 1992.

[178] **Kästner, J.**, Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6): 932–942. 2011.

[179] **Roux, B.**, The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1): 275–282. 1995.

[180] **Frenkel, D. and Smit, B.**, *Understanding molecular simulation: from algorithms to applications*. Academic Press, Amsterdam, Netherlands. 2002.

[181] **Torrie, G.M. and Valleau, J.P.**, Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2): 187–199. 1977.

[182] **Hansen, J.P. and McDonald, I.**, *Theory of Simple Liquids*. Elsevier, Amsterdam, Netherlands. 2006.

[183] **Jarzynski, C.**, Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14). 1997.

[184] **Kumar, S., Rosenberg, J., Bouzida, D., Swendsen, R., and Kollman, P.**, The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8): 1011–1021. 1992.

[185] **Hub, J.S., de Groot, B.L., and van der Spoel, D.**, g_wham A free weighted histogram analysis implementation including robust error and autocorrelation estimates. *Journal of Chemical Theory and Computation*, 6(12): 3713–3720. 2010.

[186] **Efron, B.**, Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26. 1979.

[187] **Hanwell, M.D., Curtis, D.E., Lonie, D.C., Vandermeersch, T., Zurek, E., and Hutchison, G.R.**, Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics*, 4(1): 1–17. 2012.

[188] **BitBucket:** Tampere University of Technology, Biophysics. `https://bitbucket.org/account/signin/?next=/biophys-tut/topologies`, accessed: 2013-11-01.

[189] **Van Rossum, G.**, Python programming language. In *USENIX Annual Technical Conference*. 2007.

[190] **Stone, J.E.**, *An efficient library for parallel ray tracing and animation*. Master's thesis, University of Missouri, Rolla. 1998.

[191] **Banerji, S., Hide, B.R., James, J.R., Noble, M.E., and Jackson, D.G.**, Distinctive properties of the hyaluronan-binding domain in the lymphatic endothelial receptor Lyve-1 and their implications for receptor function. *Journal of Biological Chemistry*, 285(14): 10724–10735. 2010.

[192] **Guvench, O., Mallajosyula, S.S., Raman, E.P., Hatcher, E., Vanommeslaeghe, K., Foster, T.J., Jamison, F.W., and MacKerell Jr, A.D.**, CHARMM additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate–protein modeling. *Journal of Chemical Theory and Computation*, 7(10): 3162–3180. 2011.